UNIVERSIDAD
DE MÁLAGA

LENGUAJES Y
CIENCIAS DE LA
COMPUTACIÓN
UNIVERSIDAD DE MÁLAGA

TESIS DOCTORAL

Application of Semantics to Solve Problems in Life Sciences

E.T.S.I. Informática
R.D. 99/2011

Autor

María Jesús García Godoy

Directores

Dr. José F. Aldana Montes
Departamento
Lenguajes y Ciencias de la Computación
Universidad de Málaga

Dr. Ismael Navas Delgado
Departamento
Lenguajes y Ciencias de la Computación
Universidad de Málaga

October 2018

UNIVERSIDAD DE MÁLAGA

AUTOR: María Jesús García Godoy

http://orcid.org/0000-0003-1976-023X

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

Departamento de Lenguajes y Ciencias de la Computación
Escuela Técnica Superior de Ingeniería Informática
Universidad de Málaga

Los Dres. **José F. Aldana Montes**, Profesor Catedrático del Departamento de Lenguajes y Ciencias de la computación de la Universidad de Málaga, y Ismael Navas Delgado, Profesor Titular del departamento de Lenguajes y Ciencia de la Computación de la Universidad de Málaga,

Certifican

que, Dña. **María Jesús García Godoy**, Licenciada en Biología por la Universidad de Málaga, ha realizado en el Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga, bajo sus direcciones, el trabajo de investigación correspondiente a su Tesis Doctoral por Compendio titulada:

Application of Semantics to Solve Problems in Life Sciences

Revisado el presente trabajo, estimamos que puede ser presentado al tribunal que ha de juzgarlo. Y para que conste a efectos de lo establecido en la legislación vigente, autorizamos la presentación de esta Tesis Doctoral en la Universidad de Málaga.

En Málaga, Septiembre del 2018

Fdo: Dr. José F. Aldana Montes

Dr. Ismael Navas Delgado

Acknowledgements

First, I would like to thank my supervisors Prof. José F. Aldana Montes and Prof. Ismael Navas-Delgado for giving me the opportunity of doing a PhD thesis despite I'm from a different area.
I also wish to thank everyone who accepted to be part of my thesis committee, for agreeing so quickly, and making it all so easy for me. Thank you Prof. Antonio Nebro!. In addition I'd like to thank my external evaluators for their corrections that improved my manuscript.
My sincere thanks to Dr. Manuel López Ibañez, Dr. Julia Handl and Dr. Richard Allmendinger for the wonderful stay we had at the Business School at the University of Manchester, in particular Julia who I worked with and helped me with the paperwork to submit this thesis.
I would also like to thank other members of the Grupo de Ingeniería del Software de la Universidad de Málaga (GISUM) who I have shared personal experiences with them. I want to thank Lisa Huckfield for her corrections in all my papers, without her, this thesis would not be possible. I also want to thank all the staff at the Ada Byron research centre, specially to Mari Carmen Villena, Belén from the reception, Belén from the cafeteria, Rosa, Raúl, Pepe, Soledad, Jorge, Juan Carlos Valdivia, and Lola.
Thank you very much my friends that stayed close to me despite I had not time to go out with them. I would like to mention Fernando Moreno, Loli Burgueño, Trini and my dear friend Marcos Arjona that helped me a lot to bear difficult times.
I don't want to forget to thank all my family, especially my parents (mum and dad), my brother (Fernando), my in-law family (Manolo, Virginia and my two little in-law sisters, Julia and Gloria) and my cat Missi. Thank you very much my granddad because despise he motivated me to write this PhD dissertation, he could not see me as a PhD in Informatics.
And last but not least, I would like to thank above all Esteban López Camacho. You're smart and can work wherever you want!. Thank you very much for teaching me informatics and be patient with me.

Contents

Resumen

La cantidad de información que se genera en la Web se ha incrementado en los últimos años. La mayor parte de esta información se encuentra accesible en texto, ya que el principal usuario de la Web es el ser humano. Sin embargo, a pesar de todos los avances producidos en el área del procesamiento del lenguaje natural, los ordenadores tienen problemas para procesar esta información textual.
Sin embargo, existen dominios de aplicación en los que se están publicando grandes cantidades de información disponible como datos estructurados. Así, en el campo de las Ciencias de la Vida se ha generado una enorme cantidad de datos estructurales durante la última década. El análisis de estos datos es de vital importancia no sólo para el avance de la ciencia, sino para producir avances en el ámbito de la salud. Sin embargo, estos datos están localizados en diferentes repositorios y almacenados en diferentes formatos que hacen difícil su integración. En este contexto, el paradigma de los Datos Vinculados ha emergido como un conjunto de buenas prácticas para conectar, compartir y exponer datos y conocimiento. Esta tecnología incluye la aplicación de algunos estándares propuestos por la comunidad W3C tales como HTTP URIs, los estándares RDF y OWL y el lenguaje de consulta SPARQL.
La comunidad de los Datos Vinculados abiertos han fomentado la publicación de conjuntos de datos entrelazados. El número de conjuntos de datos y triplas RDF publicados en LOD se ha incrementado en esta última década. En 2007, los 12 conjuntos de datos que formaban la nube de Datos Vinculados contenía más de 2 billones de triplas y 2 millones de documentos RDF. En el año 2013, la nube contenía 2.289 repositorios y más de 11 billones de triplas atendiendo a las estadísticas oficiales publicadas. Actualmente el número de repositorios corresponde a 2.973 y el número de triplas equivale a 140 billones. Las Ciencias de la Vida fue uno de los primeros campos en adoptar la tecnología de los Datos Vinculados. Esta adopción se ha materializado en una gran cantidad de datos en este campo que corresponde a un 11,05% del total de la nube de Datos Vinculados. No obstante, a pesar de los esfuerzos en publicar datos usando estas tecnologías, existe cierta escasez de aplicaciones que recuperen información, proporcionen nuevas perspectivas a los Datos Vinculados ya publicados o generen nuevas asociaciones en RDF que pueden resultar útiles. Dado este contexto, se han identificado un conjunto de limitaciones en la tecnología de los Datos Vinculados en el dominio de las Ciencias de la Vida:

- La disponibilidad del conjunto de datos. Cuando se habla de disponibilidad, se hace referencia a que existen repositorios RDF que han dejado de funcionar debido a la ausencia de mantenimiento por parte de los proveedores de datos. Actualmente, nos encontramos con un clara diferenciación en estos repositorios que no suelen aplicar técnicas que aseguren la disponibilidad de los mismos.

- La heterogeneidad semántica. Este problema se refiere al hecho de que cada repositorio RDF presenta diferentes vocabularios, URIs, modelos de datos, etc. Esto complica la integración de estos datos y su reutilización.

- Una curva de aprendizaje acusada que se refiere a las dificultades que los usuarios presentan para aprender esta tecnología.
- Publicación de modelos de datos.

El primer problema derivado del uso de esta tecnología se refiere a la disponibilidad de los repositorios RDF. En la categoría de las Ciencias de la Vida, existen varios ejemplos como Bio2RDF que presentan este problema. Bio2RDF es un proyecto que se constituyó en el año 2008 que integra múltiples fuentes de datos RDF que pueden ser consultados. No obstante, muchos de estos *endpoints* han dejado de funcionar a excepción de Bio2RDF *endpoint* que sigue aún operativo. Para solventar el problema, los desarrolladores de Bio2RDF han proporcionado una página basada en JavaScript que contiene volcados de datos que los usuarios pueden descargar. Sin embargo, la página no es fácil de analizar y los datos no se actualizan frecuentemente. Otro ejemplo de este problema es BioPortal que proporciona un *endpoint* siempre disponible pero su información no se ha actualizado en años. Sin embargo, un ejemplo de proyecto de éxito que ha superado tal problemática es la plataforma RDF EBI cuyos servicios han estado siempre disponibles desde su lanzamiento, en el año 2014. Por lo que la tendencia para mantener la evolución de esta tecnología requiere que sean los propietarios de los datos los que asuman su publicación como Datos Vinculados.
El segundo problema se refiere a la heterogeneidad semántica existente en los repositorios RDF en el campo de Ciencias de la Vida. Un ejemplo que refleja este tipo de problema es PhLeGrA. Esta plataforma, publicada en el año 2017, tiene como objetivo integrar cuatro fuentes de datos RDF sobre farmacología mediante la federación de consultas en SPARQL con el objetivo de descubrir nuevas asociaciones farmacológicas implícitas en el grafo RDF. En este proyecto, los autores detectaron que el mismo fármaco puede ser representado varias veces por diferentes URIs. Los autores reflejaron la necesidad de un proceso de reconciliación de URIs que no es posible solucionar mediante el uso de la federación de consultas SPARQL. Una solución propuesta se basó en el hecho de centralizar datos usando un esquema y un servicio común, sin embargo, esta estrategia se aleja de los principios propuestos por la tecnología de los Datos Vinculados.
El tercer problema hace referencia a la acusada curva de apredizaje por parte de los usuarios finales para entender la tecnología los Datos Vinculados. Esto provca que el uso de los Datos Vinculados se limite a los desarrolladores de aplicaciones, reduciendo el impacto de esta tecnología. Es necesario, por tanto, producir soluciones que acerquen esta tecnología a los usuarios finales, ampliando su impacto a medio-largo plazo.
Otro problema que se detectó en la aplicación de la tecnología de los Datos Vinculados fue la ausencia de los modelos semánticos en repositorios RDF. La construcción de repositorios RDF está guiada por el uso de un modelo semántico que proporciona todos los elementos que representan un grafo RDF. Este modelo proporciona todas las clases y relaciones que son usadas para la descripción de los datos. Las consultas SPARQL usarán estos elementos para la extracción de datos. El desconocimiento del modelo de datos subyacente dificulta el diseño de consultas, que en el caso de los Datos Vinculados Abiertos no realiza el propietario de los datos. La mayoría de las técnicas presentadas en la literatura se refieren a vocabularios y patrones que son aplicados a un conjunto reducido de repositorios.
El núcleo de los Datos Vinculados, las ontologías, también han tenido numerosas aplicaciones en el campo de la Ciencias de la Vida y de la Salud. El estándar del W3C para ontologías (OWL), se basa en las lógicas de descripciones. Esto proporciona a las ontologías y los datos (instancias) la capacidad de aplicar técnicas de razonamiento. Esta característica es aprovechada por la familia de aplicaciones software conocidas como razonadores. Sin embargo, y pese al amplio uso de las ontologías en las Ciencias de la Vida, esta capacidad de razonamiento no se aprovecha normalmente. Visto el razonamiento como un proceso que incluye la clasificación, su aplicación en salud sería

muy relevante. Así, existen, por ejemplo, aproximaciones a la representación con ontologías de terminologías médicas como ICD-10 pero no sacan provecho de esta capacidad de razonamiento (clasificación).

Objetivos

En las Ciencias de la Vida, los Datos Vinculados son considerados una pieza clave para interconectar información procedente de diferentes fuentes mediante el uso de estándares para facilitar el proceso de publicación, compartición y reutilización de datos. Atendiendo las motivaciones descritas en la sección anterior, se presentan los siguientes objetivos:

- Promover el uso de los Datos Vinculados en el ámbito de las Ciencias de la Vida.
- Facilitar el diseño de consultas SPARQL mediante el descubrimiento del modelo subyacente en los repositorios RDF.
- Demostrar la viabilidad en el uso de razonamiento en problemas de clasificación en el ámbito de la salud.
- Crear un entorno colaborativo que facilite el consumo de Datos Vinculados por usuarios finales, que promueva el uso de este entorno para propiciar el uso de esta tecnología por parte de estos usuarios. En este entorno se diseñarán y publicarán consultas federadas que recuperen información de más de un repositorio para promover su uso entre la comunidad de usuarios.
- Desarrollar un algoritmo que, de forma automática, permita descubrir el modelo semántico en OWL de un repositorio RDF. Este método se basará en un conjunto de consultas SPARQL que exploren la estructura del grafo RDF. Como resultado, se pretende poder extraer el modelo semántico para ayudar a usuarios a diseñar nuevas consultas en SPARQL.
- Desarrollar una representación en OWL de ICD-10-CM llamada Dione que ofrezca una metodología automática para la clasificación de enfermedades de pacientes en un contexto médico. Para ello, se aprovechará la existencia *mappings* SNOMED CT e ICD-10-CM
- Se validará esta ontología utilizando un razonador OWL, previamente poblada con datos (instancias) de casos clínicos de expedientes de pacientes de un centro hospitalario.

Fases

Para llevar a cabo los objetivos mencionados para cada línea de investigación, se completaron las siguientes fases:

- Análisis del estado del arte actual sobre estudios que intentan aplicar la tecnología de los Datos Vinculados al dominio de las Ciencias de la Vida. En esta exploración se pretendía elaborar la hipótesis de que la mayoría de las aplicaciones disponibles no llegan a usuarios con un perfil biológico dada la curva de aprendizaje tan acusada para entender este tipo de tecnología. Asimismo, se debía confirmar la ausencia de técnicas automáticas para la extracción del modelo semántico implícito de repositorios RDF para facilitar la construcción de consultas SPARQL. Adicionalmente, como caso de uso real que demostrara las ventajas de esta tecnología, se investigó el uso de ontologías para la clasificación de enfermedades.

- Diseño e implementación de un espacio colaborativo para el fomento del uso de Datos Vinculados en las Ciencias de la Vida. El desarrollo no se limita al software sino a que se aborda la creación y registro de una semilla de consultas federadas y no federadas. Además, se ofrecerá la evaluación de las consultas publicadas. La validación del funcionamiento de las consultas para asegurar su éxito a medio-largo plazo.

- Diseño e implementación de una técnica automática para el descubrimiento del modelo semántico implícito en cualquier repositorio RDF. Para la validación de este algoritmo, se planifica el uso varios casos de uso tales como LinkedGeoData, kpath, ReprOlive y Biomodels.

- Diseño e implementación una representación en OWL 2 cuyas inclusiones y exclusiones aprovechen los *mappings* SNOMED CT/ICD-10-CM generados por UMLS. Esta fase incluye la validación con un razonador. Finalmente, se incluye la validación con casos de uso reales para mostrar su aplicabilidad como sistema de clasificación de patologías.

- Publicación de resultados de las distintas soluciones software en diferentes revistas JCR y congresos de prestigio.

Contribuciones

Finalmente, las principales contribuciones de esta tesis son las siguientes:

- Propuesta e implementación de Bioqueries[1] como repositorio central de consultas en SPARQL federadas y no federadas en el campo de las Ciencias de la Vida. Estas consultas pueden ser diseñadas mediante una exploración visual previa de las relaciones entre recursos de un repositorio RDF. En Bioqueries, los usuarios también pueden crear, compartir o ejecutar consultas SPARQL documentadas en lenguaje natural. Bioqueries contribuye a aproximar expertos en Ciencias de la Vida a la tecnología de los Datos Vinculados.

- Propuesta e implementación de un sistema[2] para la extracción del modelo semántico de una base de datos RDF accesible a través de un SPARQL *endpoint*. El algoritmo desarrollado soluciona el problema mediante el descubrimiento de estos modelos semánticos. El código del algoritmo está disponible para la comunidad de desarrolladores con el objetivo de mejorarlo y/o extenderlo. Asimismo, se ha proporcionado un portal mediante el cual un usuario final puede ejecutar el algoritmo y extraer el modelo semántico de cualquier repositorio.

- Propuesta y desarrollo de la primera representación en OWL 2 (Dione)[3], lógicamente consistente, que modela el ICD-10-CM. Dione puede ser usado como un sistema de clasificación de enfermedades ya que los axiomas definen inclusiones del ICD-10-CM a través de los *mappings* SNOMED CT/ICD-10-CM. Dione actualmente presenta 391.669 clases, 391.720 axiomas anotaciones y 11,795 axiomas de tipo *owl:equivalentClass* que se han construido mediante 104,646 relaciones extraídas de los *mappings* SNOMED CT/ICD-10-CM de UMLS y Bioportal. La validación de la ontología se consolidó mediante su razonamiento con Elk y su aplicación a casos de usos reales. Estos resultados muestran que Dione es una herramienta prometedora para la clasificación de patologías y ayuda en el diagnóstico a especialistas de la salud.

[1] `http://bioqueries.uma.es`
[2] `http://khaos.uma.es/oee/`
[3] `http://khaos.uma.es/dione/`

Trabajos publicados

El trabajo realizado en esta tesis ha dado lugar a varias publicaciones y divulgaciones científicas. Específicamente, tres artículos han sido publicados en revistas indexadas en el *Journal of Citation Report* (JCR) del *Institute of Scientific Information*. Asimismo, otros cuatro artículos han sido publicados en congresos. Dos de ellos se publicaron en congresos internacionales y otros dos en un congreso nacional.

En el grupo de revistas JCR, se ha excluído el artículo titulado "kpath: integration of metabolic pathway linked data" dado que fue un trabajo colaborativo entre nuestro grupo de investigación Khaos y el grupo de investigación ProCel, perteneciente al departamento de Bioquímica y Biología Molecular de la Universidad de Málaga [1]. En el grupo de artículos de congreso, se excluyeron dos publicaciones que se publicaron en las actas de las JISBD (Jornadas de Ingeniería del Software y Bases de Datos) debido a que el primer artículo tiene como finalidad la difusión del trabajo realizado y el segundo consiste en un trabajo emergente redactado en español cuyo objetivo es, mediante un proceso automático, poblar el modelo OWL de ICD-10-CM (Dione) para la obtención de una versión completa de éste.

A continuación, se resumen los artículos que avalan esta tesis. En el primer artículo se presenta una versión preliminar de Bioqueries. En el segundo artículo se describe una versión mejorada de Bioqueries con nuevas características. En el tercer artículo, a partir de los problemas derivados del proyecto de Bioqueries se implementó una técnica basada en la extracción del modelo semántico de repositorios RDF. En el cuarto artículo se presenta Dione que es la primera representación en OWL del ICD-10-CM lógicamente consistente y que se ha aplicado a casos de uso reales.

Bioqueries: a social community sharing experiences while querying biological linked data

M. J. García-Godoy, I. Navas-Delgado, and J. Aldana-Montes. "Bioqueries: A Social Community Sharing Experiences While Querying Biological Linked Data". *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*. SWAT4LS '11. London, United Kingdom: ACM, 2012, pp. 24–31. ISBN: 978-1-4503-1076-5. DOI: 10.1145/2166896.2166906. URL: http://doi.acm.org/10.1145/2166896.2166906

En este trabajo se presentó un estudio preliminar de Bioqueries, una comunidad social en la cual usarios pueden diseñar, documentar y ejecutar consultas SPARQL (no federadas) con el objetivo de fomentar el consumo de Datos Vinculados en el campo de las Ciencias de la Vida. Esta herramienta fue diseñada para varios perfiles de usuarios: perfile bioinformático y perfil biológico. Los usuarios con el primer perfil son capaces de diseñar y generar consultas SPARQL y los usuarios del segundo perfil pueden hacer uso de estas consultas y recuperar información en RDF mediante su ejecución. Esta herramienta se construyó como una aplicación web en el formato común de una wiki. Cuando los usuarios se registran en la aplicación, obtienen la capacidad de crear sus propias consultas SPARQL que serán ejecutadas en una lista de *endpoints* públicos externos con datos vinculados. Esta colección de *endpoints* habrá sido añadida con anterioridad por los administradores de la plataforma. Cada consulta SPARQL puede ser documentada y debe estar especificada como una sentencia parametrizada y legible para el ser humano. De esta manera, cuando los usuarios deseen ejecutar la consulta, se encontrarán con un formulario fácil de rellenar en vez de el código de la consulta SPARQL cuando deseen consultar una base de datos externa. El resultado de cada consulta se presenta en formato tabla y puede ser exportado en varios formatos comunes de fichero de datos vinculados. A cada usuario se le permite editar y ejecutar sus propias consultas, y además

[1] http://www.bmbq.uma.es/procel/

tienen la posibilidad de hacerlas públicas para que puedan ser ejecutadas y compartidas con el resto de la comunidad.
Para fomentar que nuevos usuarios comenzaran a usar nuestra plataforma, Bioqueries fue poblada en un inicio con un conjunto de 116 consultas públicas prediseñadas. Todas ellas fueron categorizadas según la naturaleza biológica de los datos consultados. De esta manera, los usuarios pueden examinar y buscar consultas en la plataforma filtrando por sus categorías, *endpoints* consultados o usando una simple búsqueda de texto.
Posteriormente, una mejora de la herramienta incorporó más características atendiendo a las primeras impresiones de los usuarios interesados en consumir Datos Vinculados.

Sharing and executing linked data queries in a collaborative environment

M. J. García Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Sharing and executing linked data queries in a collaborative environment". *Bioinformatics* 29.13 (2013), pp. 1663–1670. DOI: 10.1093/bioinformatics/btt192
Factor de Impacto (2013): 4,621. Q1 (2/57) en la categoría de Matemáticas y Biología Computacional.
En este trabajo se presentó una versión mejorada de Bioqueries atendiendo a las sugerencias que se realizaron en un estudio preliminar. Una nueva característica que se incorporó fue la implementación de consultas SPARQL federadas que recuperan información de más de una fuente de datos y pueden generar respuestas a consultas complejas. Este tipo de consultas realizan una o más llamadas de protocolo SPARQL en una consulta, mediante la separación de ésta en diferentes subconsultas que son ejecutadas y sus resultados son integrados. Por otro lado, para favorecer la exploración del grafo RDF así como la construcción y diseño de consultas se incorporó Relfinder. Este software se utiliza para visualizar relaciones entre dos o más nodos del grafo RDF y encontrar nuevas asociaciones.
En ocasiones, la disponibilidad de los *endpoints* externos incluidos se puede encontrar limitada temporalmente a causa de mantenimiento o fallo de sus servidores, problemas de red u otros problemas diversos. Ya que esta situación está fuera de control de los administradores de Bioqueries, se incluyó un panel de información con el estado actualizado de los *endpoints*. Adicionalmente, antes de que un usuario intente ejecutar una consulta, Bioqueries comprueba automáticamente el estado del correspondiente *endpoint* y desactiva la posibilidad de ejecutar una consulta cuando el *endpoint* se encuentre no disponible.
Asimismo, se realizó un estudio de usabilidad de la plataforma desde que se hizo pública. El número de consultas almacenadas fueron 215 (siendo un 5,6% realizadas por usuarios externos) y el número de usuarios 230. Se realizó un estudio basado en la puntuación SUS [2] para dos tipos de perfiles de usuarios, biológico y bioinformático, cuyos valores correspondieron a 78,3 y 79,6, respectivamente.

Re-constructing Hidden Semantic Data Models by Querying SPARQL Endpoints

M. J. García-Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Reconstructing Hidden Semantic Data Models by Querying SPARQL Endpoints". *Database and Expert Systems Applications: 27th International Conference, DEXA 2016, Porto, Portugal, September 5-8, 2016, Proceedings, Part I.* ed. by S. Hartmann and H. Ma. Cham: Springer International Publishing, 2016, pp. 405–415. ISBN: 978-3-319-44403-1. DOI: 10.1007/978-3-319-44403-1_25

[2] https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html

Este trabajo se presentó en el congreso del DEXA 2016 celebrado en Oporto, Portugal en Septiembre de 2016. Este trabajo surgió de la problemática de diseñar consultas SPARQL para Bioqueries. Idealmente, el modelo semántico subyacente de los datos debería estar documentado utilizando el vocabulario estándar VoID o, de forma más simple, utilizando HTML. Sin embargo, esto no siempre es así. Atendiendo a esta problemática, se ha desarrollado una herramienta que, automáticamente, infiere la estructura RDF implícita de un *endpoint*.
El método que esta herramienta utiliza se basa en un conjunto se consultas SPARQL que son ejecutadas para inferir la estructura RDF implícita de un endpoint. En una primera fase, las consultas son ejecutadas para recuperar el conjunto de clases y de propiedades presentes en el modelo del repositorio RDF. Posteriormente, dos consultas serán ejecutadas para cada propiedad para obtener su dominio y rango. Atendiendo al rango de la propiedad, es posible dividir la lista obtenida de propiedades en dos subconjuntos: propiedades de objeto (el rango está compuesto por clases) y propiedades de tipos datos (cuando el rango está compuesto por tipos de datos). Finalmente, el resultado de estas consultas se combinan para generar una ontología OWL que representa un modelo semántico de los datos.
Esta herramienta se implementó en Java y su código es disponible en Github. También se implementó una aplicación web que proporciona todas las funcionalidades de la herramienta, de tal manera que, los usuarios no tienen que compilar y ejecutar el código. Los resultado obtenidos pueden visualizarse utilizando esta web o, bien, pueden ser exportados y descargados en un fichero OWL.
Esta herramienta se testéo utilizando tres casos de uso: dos *endpoints* SPARQL desarrollados por nuestro grupo de investigación (Kpath y ReprOlive), así como un *endpoint* SPARQL popular entre la comunidad de los Datos Vinculados: LinkedGeoData. Estos tres ejemplos se seleccionaron atendiendo al tamaño de los datos y su complejidad para medir la calidad de la ontología generada. El código del algoritmo está disponible para la comunidad de desarrolladores y para usuarios finales. Adicionalmente, se implementó un servicio Web que proporciona resultados de los casos de uso aplicados y ejecutar el algoritmo para cualquier repositorio RDF.

Dione: An OWL representation of ICD-10-CM for classifying patients' diseases

M. del Mar Roldán-García, M. J. García-Godoy, and J. F. Aldana-Montes. "Dione: An OWL representation of ICD-10-CM for classifying patients' diseases". *Journal of Biomedical Semantics* 7.1 (2016). DOI: 10.1186/s13326-016-0105-x
Factor de Impacto (2016): 1,845. Q2 (18/57) en la categoría de Matemáticas y Biología Computacional. Este artículo se publicó en el Journal of of Biomedical Semantics. Este artículo se focaliza en la importancia de formalizar terminologías biomédicas como ICD-10-CM utilizando OWL. Atendiendo a la literatura revisada, han habido varias investigaciones que no han podido modelar el ICD-10-CM como representación OWL. Uno de los modelos propuestos captura la jerarquía del ICD-10 y modela las inclusiones y exclusiones con un componente OWL-full. Aunque los autores del modelo sugirieron que parte de la ontología se expresa mediante el uso del perfil OWL-DL y un componente OWL full para aquellas propiedades que exceden la expresividad OWL-DL, el modelo propuesto no se validó con un razonador OWL. Esto se explica ya que el modelo, que tiene un componente OWL-full, no puede ser validado por un razonador OWL. Por tanto, atendiendo a la literatura revisada y a la ausencia de modelos OWL para ICD-10, se propuso Dione, que corresponde al primer modelo OWL, lógicamente consistente, cuyos axiomas que definen las inclusiones y exclusiones ICD-10-CM están basadas en los *mappings* entre SNOMED CT e ICD-10-CM. Estos *mappings* oficiales, se han extraído de UMLS. Los axiomas extraídos de los *mappings* oficiales completaron Dione en un 93% lo que significa que el 93% de las categorías ICD-10-CM presentan al menos un axioma que la define. Por ello, para generar una versión más completa del

modelo en OWL del ICD-10-CM, se utilizaron otros *mappings* procedentes de BioPortal los cuales fueron validados de forma semi-automática mediante el algoritmo LOOM y la colaboración de los usuarios de la comunidad científica. Dione actualmente contiene 391.669 clases, 391.720 axiomas de anotaciones y 11,795 axiomas de tipo *owl:equivalenClass* extraídos de los *mappings* SNOMED CT/ICD-10-CM utilizando como elemento conjuntivo *owl:intersectionOf* y *owl:someValuesFrom*. La ontología resultante ha sido clasificada mediante el uso del razonador ElK. Asimismo, para demostrar la utilidad de Dione, se han usado anotaciones SNOMED CT obtenidas manualmente procedentes de expedientes clínicos de pacientes del Hospital Virgen de la Victoria (Málaga). Esta información fue formalizada mediante instancias y clasificadas por el razonador Elk. Los resultados de la clasificación en estos casos de uso muestran que Dione podría ser un prometedor modelo semántico en OWL para apoyar a los especialistas médicos en la clasificación de patologías.

Conclusiones y Trabajos Futuros

La cantidad de datos generados en la Web ha crecido de forma exponencial en las últimas décadas, concretamente, en el campo de las Ciencias de la Vida. Esta cantidad de datos ha propiciado la necesidad de incluir semántica procesable por las máquinas. La tecnología de los Datos Vinculados emergió en el año 2008 y se define como un conjunto de prácticas recomendadas para compartir, exponer y conectar datos mediante el uso de estándares tales como URIs, RDF y OWL. Esta tecnología ha sido respaldada por la comunidad W3C extendiéndose a diferentes categorías de datos, por ejemplo categorías de redes sociales, geografía, lingüística, dominios cruzados, medios de comunicación y Ciencias de la Vida.

Las Ciencias de la Vida fue uno de los primeros dominios en los que se adoptó la tecnología de los Datos Vinculados. La incorporación de esta tecnología ha sido propiciada por la cantidad de datos generados debido a los avances tecnológicos en este ámbito. Sin embargo, esta tecnología ha presentado ciertos problemas relacionados con: 1) la disponibilidad de los repositorios RDF, 2) la heterogeneidad de la semántica utilizada y 3) la curva de aprendizaje por parte de los especialistas en el campo de las Ciencias de la Vida.

Atendiendo a los problemas especificados, esta tesis propone varias soluciones que combinan el estudio de cada problemática con el desarrollo de software. Bioqueries supuso la primera aportación de esta tesis, proporcionando un entorno colaborativo para la comunidad de las Ciencias de la Vida. Este entorno ha tenido un éxito relativo que se refleja en un número elevado de usuarios. Sin embargo, una de las carencias detectadas es la baja actividad de esta comunidad de usuarios. Esto se refleja en el bajo índice de publicación de consultas (sólo el 5,6% de las consultas). Sin embargo, el uso de estas consultas sí es mayor. Esto se debe en parte al esfuerzo en el diseño de un conjunto elevado de consultas y al desarrollo de herramientas de soporte como la detección de repositorios que no están funcionando. Como primera aproximación a la baja actividad de publicación de consultas, se ha abordado el problema del conocimiento del modelo subyacente en estos repositorios. En un escenario ideal, este modelo debería estar documentado sin embargo, en la práctica, los desarrolladores no suelen proporcionarlo. Para resolver este problema, se ha diseñado e implementado un algoritmo que permite la reconstrucción automática del modelo semántico del repositorio RDF. El resultado fue una herramienta que está disponible para la comunidad científica para su extensión y mejora, así como un servicio Web en el que el algoritmo puede ser ejecutado y los resultados mostrados. El algoritmo se testeó realizando un conjunto de experimentos en los que se llevaron a cabo mediante el uso de repositorios como LinkedGeoData, Biomodels, ReprOlive y kpath. Sin embargo, a pesar de que el algoritmo recuperó el modelo semántico del repostortio, tal recuperación fue parcial. Un ejemplo de este problema es que la relación entre clases no se obtiene si no se especifica en el repositorio. Es por ello, que se trabaja actualmente en una segunda versión basada en obtener relaciones mediante dos estrategias: 1) aplicar un método basado en el

alineamiento de ontologías y completar las relaciones no recuperadas y 2) extraer el conjunto de instancias por clase y calcular si alguna de ellas presentan alguna relación semántica. Este proceso puede curarse manualmente si algunas de las relaciones extraídas son falsos positivos. Además, se plantea como trabajo post-doctoral la integración con Bioqueries. Por lo que se espera aumentar el ratio de participación de sus usuarios. Además, una posible solución para el problema que presenta Bioqueries es el lanzamiento de nuevas características que ayuden a los usuarios finales a construir las consultas en SPARQL de forma semi-automática y la incorporación de consultas complejas consolidando Bioqueries como repositorio universal de consultas en las Ciencias de la Vida.
La ontologías son parte del conjunto estándares que la tecnología de los Datos Vinculados aconseja utilizar como esquema de datos. Esto ha permitido que las ontologías se usen ampliamente en los campos de las Ciencias de la Vida y de la Salud. Una ontología OWL permite, de acuerdo a los diferentes perfiles OWL, ser razonada por razonadores OWL. Esto ha tenido una gran implicación en Biomedicina ya que se pueden usar como sistemas de clasificación para patologías utilizando terminologías biomédicas. Un importante reto fue el modelado del ICD-10 en OWL. Atendiendo a la literatura, han surgido mucho estudios basados en modelar el ICD-CM en OWL sin embargo, todos muestran ciertos defectos en cuanto a la implementación de un modelo OWL que sea lógicamente consistente. En esta tesis se presenta Dione, la primera representación en OWL del ICD-10-CM, que corresponde a la última versión de ICD. Para materializar esto, se implementó un proceso automático para la generación de un árbol jerárquico con clases de enfermedades del ICD-10-CM y sus axiomas modelados como *owl:equivalentClass*. El resultado de este modelo fue un modelo en OWL que puede ser utilizado por un razonador. Dione está completa al 93,3% lo que significa que el porcentaje de clases restantes no presentan axiomas. Por tanto, el siguiente paso fue la implementación de una técnica que permite poblar Dione con axiomas extraídos de *mappings* de ICD-10-CM y una ontología diana de BioPortal. Como caso de uso, se ha usado ORDO, una ontología de enfermedades raras, genes y otras características que tiene clases equivalentes en Dione. El resultado fue una versión más completa de Dione con la adición de nuevos axiomas que definen las clases ICD-10-CM mapeadas con las de ORDO. De acuerdo con los resultados previos, se está planificando replicar el mismo experimento para todas las ontologías mapeadas con el ICD-10-CM para proporcionar la versión más completa de ICD jamás modelada.

Chapter 1

Introduction

The amount of information on the World Wide Web has enormously increased in the last few years. On the World Wide Web, machines store, organize, request, route, receive and display all this information [5]. However, computers struggle with the processing of this information despite all the advances performed in the area of Natural Language Processing (NLP).
In this context, there are domains of application in which large amounts of data have been produced. For example, in the Life Sciences domain, the amounts of data, which have been produced at unprecedented rate are enormous, specially, during the last decade. The analysis of these data is very promising to solve problems in areas such as Life Sciences and Health Care. However, these data are located in different repositories and stored in different formats that make it difficult to integrate. In this context is where the paradigm of "Linked Data" has strongly emerged as a set of good practices with the objective of connecting, sharing, and exposing data and knowledge. The underlying technology that Linked Data uses includes the application of some standards supported by the W3C such as the HTTP URIs, the SPARQL query language, RDF and OWL.
The Linked Data community has encouraged the publication of interlinked datasets since the year this paradigm emerged. The number of published datasets and triples have increased over the last decade. In 2007, the set of 12 datasets that formed the Open Linked Data Cloud contained >2 billions of RDF triples and 2 millions of RDF documents. In 2013, the Linked Data Cloud contained 2,289 datasets and more than 11 billions of triples according to the published statistics [6]. Currently, the number of datasets corresponds to 2,973 and the number of triples are more than 140 billions. Life Sciences was one of the earliest adopters of the Linked Data technology and its principles by representing, linking, publishing and querying in the Web of Data. This adoption has been materialized as a great amount of data in Life Sciences published in the Linking Open Data Cloud that corresponds to 11.05% over the total of *space* of data. However, despite of all these efforts there is a lack of major applications to retrieve information, provide new insights to the existing RDF links or even find new RDF associations that can be useful. In this context, we have identified a set of limitations of the Linked Data technology in the domain of the Life Sciences, which can be enumerated as follows:

- The availability of the data sets. With availability of these services, we are referring to RDF repositories or SPARQL endpoints that stop working due to the lack of maintenance by the data providers.
- Another problem to be considered is the semantic heterogeneity that refers to the fact that each RDF repository follows different vocabularies and URIs, data models etc.
- The steep learning curve that refers to the difficulties that many end-users (Life Science) have

to overcome to learn the Linked Data technology. For example, those end-users that have to use SPARQL language as query language to retrieve RDF information from the SPARQL endpoints.

The first problem corresponds to the availability of the RDF data sets, which has become a problem in the Linked Data technology in many categories of data. In the category of Life Sciences, there are some remarking examples of this problem such as Bio2RDF. As was mentioned, Bio2RDF is a project, launched in 2008, that integrates multiple open data sources in RDF that could be queried by SPARQL queries. However, many of these SPARQL endpoints have currently stopped working with exception of the Bio2RDF endpoint that is operative [7]. The Bio2RDF data maintainers have also provided a JavaScript-based page, which can be found in [8] that contains all RDF dumps that users can download. This page is not easily crawled by machines without a JavaScript interpreter and also the RDF dumps are not frequently updated. Another example is BioPortal that provides an SPARQL endpoint that is always available but has not been updated for years [9]. Despite all of these problems presented with this technology, there is an example of success that is the EBI RDF platform. This platform has been available since 2014 and all the SPARQL endpoints have been available and their RDF data updated from then. Therefore, for the evolution of this technology, the data providers have to take responsibility of the publication of their Linked Data.
The second problem is based on the semantic heterogeneity, which is a problem derived of the use of Linked Data in the category of Life Sciences. For example, as described in [10], one example that reflects the problem of semantic heterogeneity is PhLeGrA. This platform, published in 2017, has as objective to integrate four RDF data sources in pharmacology through SPARQL query federation, to tackle the structural heterogeneity in Life Sciences data and discover new pharmacological associations. In this project, the authors detected that the same drugs can be represented several times by different URIs in different RDF data sources. According to the authors, it is necessary a process of URIs reconciliation that can not be solved with query federation given their complexity. Therefore, in order to address the problem of URIs reconciliation, the Linked Data applications' developers use warehousing in which all data is transformed under a common schema using a uniform set of notations or shared identifier repositories. An example of this is the work developed by [11] in which RDF data sources from NCBI taxonomy, Uniprot, Kegg and Bio2RDF were stored in a common Virtuoso RDF repository whose data in RDF is retrieved by the *kpath* application. However, this solution requires a lot of maintenance efforts and updating by the developers.
The steep learning curve is another problem on the understanding of the end-users the Linked Data technology in Life Sciences. In [12], the authors stress the lack of documentation or guidelines in order to discover and reuse the information stored as RDF and interconnected to other sources and also the problems that many Life Sciences researchers have to make SPARQL queries. Taking into account all these problems, in 2013, we published the platform Bioqueries that aims to encourage the sharing of users' experience trying to take advantage of Linked Data in the Life Sciences domain [2]. Bioqueries aims to start the process towards a greater understanding of Life Sciences Linked Data sources by means of online social networks. Bioqueries opens up a way to build-up communities around a shared interest in certain biological domains to take advantage of public Linked Data. This is achieved by sharing a virtual space in a wiki-based portal for the design and execution of (federated and non-federated) SPARQL queries that are documented using natural language descriptions. However, Bioqueries currently confronts the aforementioned problem of availability as many of the SPARQL services have stopped by the maintainers with exception of some successful projects such as EBI RDF platform. The success of this project might be explained because of the its capability of data centralization as EBI is one of the largest resources on Bioinformatics in the world. However, the quality of the Open Linked Data cloud increases as much as the data providers publish their data as Linked Open Data.
Another problem that we detected was that most of the RDF repositories do not provide their

underlying semantic data models in OWL. The creation of RDF repositories is guided by using the semantic model, which provides all the elements a RDF graph represents. SPARQL queries use these elements for data extraction. The lack of the semantic models in RDF repositories makes difficult the design of SPARQL queries. The problems to extract the underlying semantic model has conducted to several works according to the literature. For example, Aemoo [13] is a tool that allows the RDF navigation through exploiting semantics and RDF links but only applicable to the Dbpedia. Linked Data summaries [14] develop and evaluate an approximate index structure summarizing graph-structured content of sources adhering to Linked Data principles, provide an algorithm for answering conjunctive queries over Linked Data on the Web exploiting the source summary. In [3], the authors tried to solve the problem by providing an algorithm that discovers part of the explicit semantic model of an RDF database through a set of simple SPARQL queries. However, despite of the last attempt, the discovery of the underlying semantic model is only partial and therefore, more studies following this research line are necessary.
The core of Linked Data, the ontologies, has been extensively applied to different domains such as Life Sciences and Health. The W3C standard for ontologies is based on description logics. This allows to apply reasoning techniques to the ontologies and data (instances). This can be taken advantage of software applications known as reasoners. However, despite the wide use of ontologies in Life Sciences, the ability of reasoning is not being exploited. Given the ontology reasoning is a process that includes classification, its application in Health would be very relevant. Therefore, there are approaches based on the representation in OWL of medical terminologies such as ICD-10 that are not taken advantage of the ontology reasoning (classification).

1.1 Objectives and Phases

In Life Sciences, the Linked Data plays a key role to interconnect data from different sources through the use of standards to ease the procedure to publish, share and reuse this data. According to the set of problems previously described in Introduction, we present the following objectives:

- Increase the use of Linked Data in the domain of the Life Sciences.
- Facilitate the design of SPARQL queries through the reconstruction of the hidden semantic model in RDF repositories.
- Show the viability in the use of reasoning in classification problems in the domain of Health.
- Create a collaborative environment that eases the consumption of Linked Data by end-users and increase its use to bring closer this technology to Life Sciences specialists or Linked Data developers. In this environment, federated SPARQL queries will be designed to retrieve information from more than one RDF repository in order to increase their use amongst the users'community.
- Develop an approach that automatically reconstruct the hidden semantic model in OWL behind an RDF database. This is based on a set of SPARQL queries to explore the structure of the RDF graph. As a result, the application is able to partially discover the underlying data model to help users in designing new SPARQL queries.
- Develop an OWL representation of the ICD-10-CM to help find an automated approach to classify patients' diseases in a medical context. The aim is to take into account the inclusions terms that are formalized by exploiting SNOMED CT/ICD-10-CM mappings.
- Populate this ontology with data (instances) from clinical use cases of patients from the Hospital Virgen de la Victoria (Málaga) and tested by using an OWL reasoner.

In order to carry out the aforementioned objectives for each research work, the following phases have been followed:

- Analysis of the current state-of-the-art of studies, which attempt to apply Linked Data to the Life Sciences domain. In this exploration, we tried to elaborate a hypothesis that most of the available applications are not usually targeted at users with a biological background given the steep learning curve to understand this technology.
- Design and implement a collaborative environment to increase the use of Linked Data in the domain of the Life Sciences. Its development is not only limited to the software but also tackles the creation and registry of a seed of non-federated and federated SPARQL queries. In addition, a system of query evaluation will be provided to assure the success in a medium and long term.
- Design and implement an approach for discovering the underlying semantic models in RDF databases. In order to test and verify this technique, we included some uses cases based such as LinkedGeoData [15], kpath [11], ReprOlive [16] and Biomodels [17].
- Design and implementation of an OWL 2 representation for ICD-10-CM whose inclusions and exclusions use the SNOMED CT/ICD-10-CM mappings generated by UMLS. This phase includes the validations with an OWL reasoner. Finally, some use cases were included to show the OWL 2 representation as a classification system of diseases.
- Publication of the results obtained from the different software solutions in conferences and JCR journals.

1.2 Thesis contributions

Finally, the main contributions of this thesis are the following:

- Proposal and implementation of Bioqueries [1] as a central repository of federated and non-federated SPARQL queries in Life Sciences. These queries can be designed by previous visual exploration of the resources relationships of an RDF repository. In Bioqueries, users can also create, share or execute SPARQL queries, which are documented by using natural language [1]. Bioqueries also contributes to bring experts in Life Sciences closer to the technology of Linked Data.
- Proposal and implementation of an approach [2] to solve the problem of discovering the hidden semantic model in RDF repositories [3]. The developed of this algorithm solves the problem through the discover of the semantic models. The code of this algorithm is openly available at [3] and can be used by developers to improve the tool to extract other aspects that the current approach could have missed. In addition, we have provided a portal to automatically generate the semantic models by including the URL of a given repository.
- The proposal of the first OWL representation of the ICD-10-CM (Dione) footnote[4], which is logically consistent. Dione can be used as a disease classification system of real patients' clinical records due to its axioms define ICD-10-CM inclusions by a methodology based on SNOMED CT/ICD-10-CM mappings [4]. Dione contains 391,669 classes, 391,720 entity

[1] http://bioqueries.uma.es
[2] http://khaos.uma.es/oee/
[3] http://github.com/estebanpua/ontology-endpoint-extraction
[4] http://khaos.uma.es/dione/

annotation axioms and 11,795 owl:equivalentClass axioms, which have been constructed using 104,646 relationships extracted from the SNOMED CT/ICD-10-CM mappings from UMLS and BioPortal. The validation of the ontology was performed through it reasoning with ELK and its application to real use cases. There results show that Dione is a promising tool for the diseases'classification and also supports the health specialists in the patients' diagnosis [4].

1.3 Thesis organization

This thesis has been organized as follows. The current chapter contains an introduction to the work done, presenting the motivation to carry it out, the objectives that have been sought, the phases that have been followed to achieve those objectives and the main contributions of the thesis. Chapter 2 includes a description of the concept of Linked Data and the components that are part of it, the technology of Linked Data applied to Life Sciences and its end-user tools of the area, the ontologies in the biomedicine and the lack of a disease classification system of ICD-10-CM according to the current literature. Chapter 3 contains all the published work that supports this thesis with a summary of each one of them. Finally, Chapter 4 includes the conclusions of this dissertation and the future research lines that can be opened by this study.

Chapter 2

State of the Art

In this chapter, we focused on the definition of the concept of Linked Data as an emerging technology with its applications to different areas. A description of the Linked Data technology is provided. Then, we present the impact and the uses that this technology of Linked Data has had in areas as Life Sciences. Finally, we concluded with the importance of ontologies in Life Sciences and Health Care, focusing on the lack of an OWL representation of ICD-10-CM, in the reviewed literature that classify diseases from patients' clinical records.

2.1 Linked Data: The Concept

The World Wide Web has revolutionized the way we publish, access and share information. The use of Web browsers enables to transversally navigate the space of information by hypertext links. Search engines index the Web documents and analyze their links structures to improve the search by string to users. This functionality has contributed to the rapid growth of the Web because of its flexible, open and extensible nature [18].

Despite all the benefits that the World Wide Web has provided since it emerged, traditional approaches are not enough at data level. The data that have been published on Web have different formalisms. However, most of these data lack of meaning and the links that connect HTML documents do not provide information about the semantics of the relations between entities. Therefore, it is necessary a more advanced technology to connect entities with enough expressivity.

In this context, where the space of information has increased and the web documents and data are linked each other, Linked Data has emerged. Linked Data is defined as a set of recommended practices for sharing, exposing and connecting pieces of information, data and knowledge by using URIs (Uniform Resource Identifiers) and RDF. This technology, supported by the W3C, allows the publication and exchange of information in an interoperable and reusable fashion. Linked Data has the property to be machine-readable, providing meaning to the data and making possible the interconnections of other external data sets to itself, and being able to be linked to and from external data sets [19].

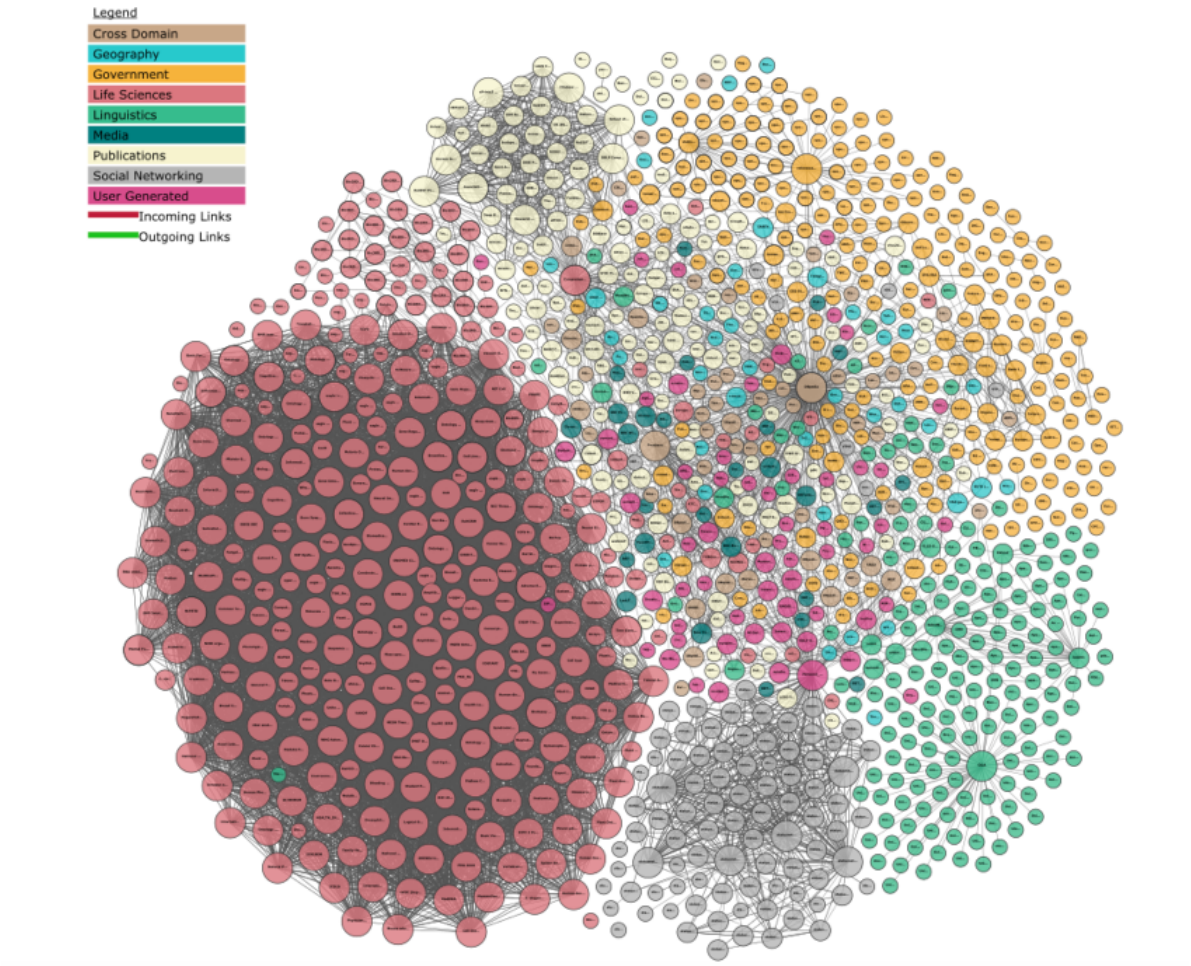

Figure 2.1: The current state of the Open Linked Data Cloud extracted from the official Web page of Open Linked Data (2018) [6].

The Linked Data principles can be summarized as Tim Berners-Lee proposed in [20]:

- Use URIs as names for things.
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful RDF information.
- Include RDF statements that link to other URIs so that they can discover related things.

The first principle refers to the use of URI references to identify not just Web documents but also real concepts or abstract things. The use of URIs is used to identify resources to abstract or real concepts.
The second principle is based on the use of HTTP URIS that in the traditional Web retrieve information by a unique identifier. However, in Linked Data, these HTTP URIs are dereferenced using the HTTP protocol to retrieve a description of the abstract or real concepts that represent.

In order to use and retrieve the interconnected data and create applications that use this technology, the standards are very important. This has led to the third principle of Linked Data that encourages the use of RDF, which is a framework for modeling information. The representation of RDF information is characterized for being minimally constrained and very flexible as the W3C specifies in [21]. This framework is described in more detail in section 2.2.

The fourth principle refers to the use of RDF links. This name is used to distinguish from hyperlinks that just connect classic Web documents. The RDF links associated to Linked Data not only connect things but also provide information about the type of association between the entities that links.

In summary, these principles that Linked Data is based on can be considered as a recipe for the community whose main efforts have focused on 1) publishing open license data sets as Open Linked Data on the Web, 2) interlinking data to other Linked Data repositories and 3) developing clients to use the Linked Data.

Since the proposal of Linked Data by Tim Berners-Lee in 2006, the Linked Open Data Cloud has exponentially increased over the last few years, creating what is called the global data space from different data sources. The topics of publishing Linked Data are very different and range from government, geography, life sciences, linguistics, etc. to data generated by users. In [22], the authors carried out an Open Linked Data statistics overview. This study reflects how Linked Data has evolved along these years (see Figure 2.2). For example, in 2011, the number of data sets was 452 and the number of RDF triples corresponded to 950 millions. In 2013, the number of datasets and RDF triples increased up to 2,289 and 11 billions, respectively. Currently, the Linked Data Cloud is formed by 9960 datasets and 149 billions of triples [6]. The current status of the Linked Data Cloud is shown in Figure 2.1.

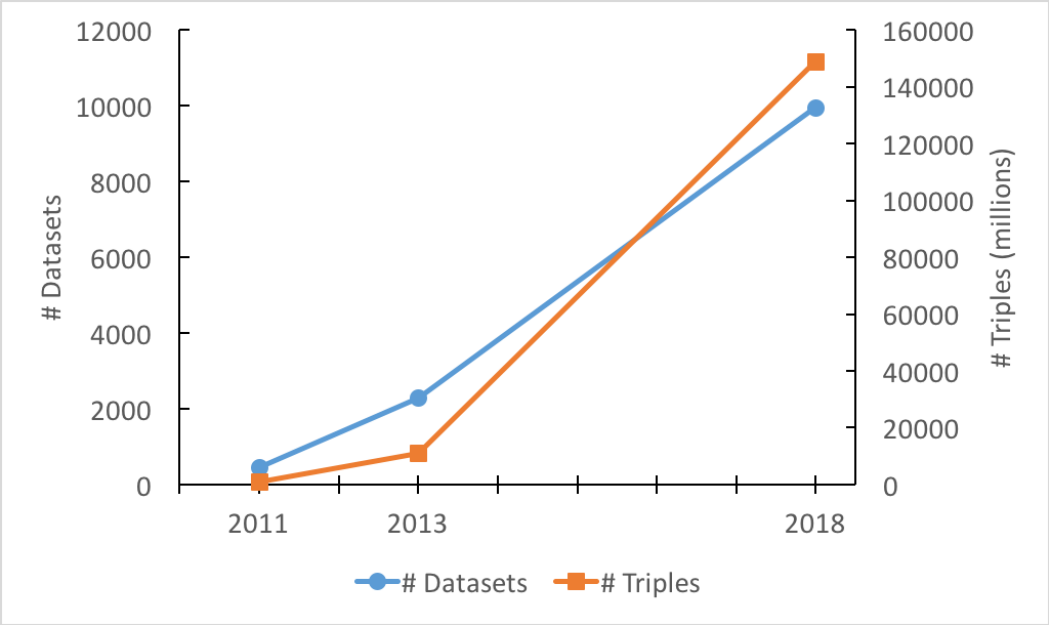

Figure 2.2: Evolution of the number of datasets and triples from 2011 and 2018. The blue and red lines represent the evolution of datasets and triples over time, respectively.

2.2 Linked Data Technology

The World Wide Web has facilitated the dissemination of information around the world. The Web structure provides what is called URL (Uniform Resource Locator) to reference Web pages and connect them with each other. However, the concept of Linked Data is based on supporting a Web of data where the data is interconnected through URIs. The main objective of Linked Data is to link data from different sources and be machine-readable. The data must have a meaning, which is explicitly defined and used to link with other external data sources. The second and third principles of Linked Data refer to the use of RDF as data model to represent this information, which is described in (see 2.2.1). Subsection 2.2.2 details the SPARQL query language used to retrieve information from endpoints that act as online query gateways to semantically annotated linked data sources. In the subsection 2.2.3, we have also included OWL that extends the RDF expressivity with additional primitives and is part of the Linked Data vocabularies together with RDF and SKOS (Simple Knowledge Organization System) .

After the concept of Linked Data emerged, interlinked RDF datasets have been created as shown in Figure 2.1, creating the Web of Data that includes cross-domain data, Life Sciences Data, government data that is fully described in subsection 2.2.4. In this context, there have been efforts to research and build applications to exploit Linked Data such as (see subsection 2.2.5) human-oriented search engines, browsers, domain-specific applications, Linked Data mash-ups, etc. Subsection 2.2.6 describes the application of Linked Data technology to the domain of Life Sciences and the role of ontologies and terminologies in Biomedicine including the importance to model the ICD-10-CM as an OWL representation.

2.2.1 RDF as Data Model

The Resource Description Framework (RDF) is a graph-oriented data model for representing information in the Web supported by the W3C consortium [23]. RDF represents the information as node-and-arc-labeled directed graphs. This data model was designed to ease the integration of information from heterogeneous data sources, being this information heterogeneous and represented in different data models. The objective was to create a standard of data model that provides interoperability between the applications that exchange machine-understandable information in the Web.

The design of RDF has met the goals of 1) providing an XML-based syntax that are the basis of the current Web technology, 2) offering a simple and open data model to represent information, 3) a language based on URIs to represent resources that can be linked to other resources and finally, 4) the semantics (or meaning) to represent the type of relations between things and with provable inference.

In RDF, a data resource is represented as a number of triples: subject, predicate and object (node-arc-node). Figure 2.3 represents an example of an RDF triple. The subject of a triple is a URI that identifies the resource that is described. In this case, the *Pathway* that corresponds with the Ubiquinone biosynthesis is the subject. The object can be a literal (a number, a string) or another resource that can be the subject of another RDF triple. The predicate, in the middle, links the subject and the object in an RDF triple and represents their relationship type. It is also represented by a URI. In this example, the predicate represents that the pathway Ubiquinone biosynthesis has a *biochemical reaction* that corresponds with the decarboxylation of 3-Octaprenyl-4-hydroxybenzoate (r04986) and is part of that pathway.

```
path00130          hasReaction          r04986
Subject            Predicate            Object
```

Figure 2.3: Abstract of a subject-predicate-object that defines an RDF triple.

The subject of an RDF triple is a URI or a blank node (a URI that is not a URI reference or literal or a URI without an intrinsic name). The predicate is always a URI. The object can be a URI (which can be the subject of another RDF triple), a literal or a blank node.

Definition 1. Formally, an RDF triple is defined as: Let U be the set of URIs, L be the set of literals, and B be the set of blank nodes. A triple $(s, p, o) \in (U \cup B) \times U \times (U \cup L \cup B)$ is called an RDF triple.

The are two types of RDF triples: Literal triples and RDF links. A literal triple refers to a triple that has a literal. A literal triple is used to describe a string, numbers or dates and can be classified as plain or type. A plain refers to a a string combined with an optional language tag. A type literal is a string combined with a datatype URI that identifies the datatype of literal. An example of typed literal is: <xsd:boolean, "true"> that refers that the specified literal in the RDF triple is a boolean and denotes the logical true that is defined in the XML schema [24]. An RDF link describes the relationship between two resources.

A set of such RDF triples is an RDF graph, which can be visualized as a node and directed arc diagram in which each triple is represented as a node-arc-node link. Figure 2.4 shows the triple that refers to the reaction that is part of the pathway defined as <http://bio2rdf.org/path:map00130> (subject). In this example, the predicate comes from the kpath endpoint [11] that integrates information about metabolism from different sources and the object and subject come from the Bio2RDF [25] that is a set of Life Sciences RDF repositories.

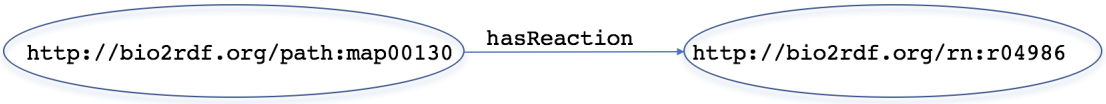

Figure 2.4: Node-Arc-Node in an RDF graph. The subject (a pathway) and object (a biochemical reaction) are connected by a predicate (*hasReaction*).

Definition 2. An RDF graph G can be defined as a set of RDF triples. Then, (s, p, o) can be represented as a direct edge-labeled graph $s \xrightarrow{p} o$.

It is worth noting that an RDF graph is not a classical graph given that a predicate (graph edge) can appear as nodes of others edges as the case of the bipartite RDF graphs [26].

An example of an RDF document using XML/RDF syntax representing the pathway map00130 corresponds with the "Ubiquinone and other terpenoid-quinone biosynthesis" pathway according to the Kegg pathway annotation (see Figure 2.5). In this figure, the pathway is represented as a URI reference <http://bio2rdf.org/path:map00130> identified by the type pathway. There are three more states that identify the related pathways to <http://bio2rdf.org/path:map00130>. This pathway is also identified by a name (<ns2:name>) and the reaction rn:r04986, which is part of.

```
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:ns2="http://khaos.uma.es/pathways/" >
  <rdf:Description rdf:about="http://bio2rdf.org/path:map00130">
    <rdf:type rdf:resource="http://khaos.uma.es/pathways/Pathway" />
    <ns2:relatedPathway rdf:resource="http://bio2rdf.org/path:map00900" />
    <ns2:relatedPathway rdf:resource="http://bio2rdf.org/path:map00400" />
    <ns2:relatedPathway rdf:resource="http://bio2rdf.org/path:map00190" />
    <ns2:name>Ubiquinone biosynthesis</ns2:name>
    <ns2:reaction rdf:resource="http://bio2rdf.org/rn:r04986" />
  </rdf:Description>
</rdf:RDF>
```

Figure 2.5: RDF representation of the map00130 pathway. The relationships <rdf:type>, <ns2:relatedPathway>, <ns2:name> and <ns2:reaction> identify the map00130 pathway.

An RDF dataset is a collection of RDF graphs and compromises one default graph, which has not a name and may be empty. It also includes zero or more named graphs. Each named graph is a pair consisting of an IRI (Internationalized Resource Identifier) or a blank node. The graphs' names must be unique.

Definition 3. A dataset can be denoted by (D, A), which is a directed graph, where D is the set of nodes (that includes URIs U, literals L and blank nodes B) and each node $D_i \in D$ denotes a dataset; A is the arc set and each arc $(D_i,\ D_j) \in A$ exists if there are at least k RDF triples $(s, p, o) \in D_i$ where s, o are two URIs in D_i and D_j, respectively. k indicates the spareness of arcs in the graph.

After emerging Linked Data, the number of available RDF datasets has increased. The W3C provides a *wiki* where the URLs of the RDF data bumps are available [27]. In this page, Linked Data consumers can find the project's name, the URL for the directory containing the RDF dump files and information about the publisher or maintainer of the RDF data. An example of this is the case of the Bio2RDF project whose main objective was to publish Life Sciences data as a biological RDF datasets [25]. In order to query the RDF data stored in the repositories, the publishers provide a service called endpoints where data can be queried with the SPARQL query language [28]. A technical definition of an endpoint is an HTTP URL, which accepts SPARQL queries and return results. The service can also return a variety of serialisations such as Turtle [29], N-triples [30], RDF/XML etc. The advantages to have these services working can be summarized as follows: 1) Access to data by Linked Data consumers, 2) Executing complex queries such as federated SPARQL queries, which retrieve information from more than one data source, 3) Retrieve information making use of different standards to be visualized how the information is interconnected. A more detailed explanation about the SPARQL query language is given in subsection 2.2.2.

Finally, we can summarize a set of advantages defined by the W3C consortium [31] as follows:

- RDF is based on XML language and support any XML datatypes.
- The RDF model composed by the subject-predicate-object structure allows to efficiently implement and store interlinked data.
- RDF provides an extensible URI vocabulary.
- The RDF model allows to make statements about any resources.
- The RDF model is based on direct edge-labeled graphs so it has the advantage to structure information using graphs.

- The RDF model can be processed in absence of more detailed information on the semantics. For example, some inferences can be logically found.

- The information on direct edge-labeled graphs can be mapped into the union with of the corresponding RDF structures.

- Information stored in RDF has advantages over the relational databases based on: 1) In RDF, the data scheme is optional, 2) The addition to new attributes to an RDF repository can be performed on the fly (in databases it requires migrations), 3) In RDF, inferences can be carried out, 4) RDF allows the detection of redundancies in case to include external data. 4) the structure of the queries allows faster joins.

2.2.2 SPARQL Query Language

SPARQL (SPARQL protocol and RDF query language) is an RDF query language proposed by the W3C consortium in order to retrieve information from RDF triple stores and manipulate it. The current version of SPARQL query language is 1.1 [32] and its official release was in 2013. The previous SPARQL query version was 1.0 and was officially proposed in 2008 [33].
As the authors Arenas *et al.* describe in [34], SPARQL queries are composed by three parts: 1) the matching pattern part that includes several interesting features of pattern matching of graphs such as optional parts, union of patterns, nesting, filtering values of possible matchings, and also the possibility of choosing the data source to be matched by a pattern; 2) the solution modifiers that once that the matching pattern has been computed allow to modify the values by applying classical operators like projection, distinct, order and limit; 3) the output from the SPARQL query that can be of different types as yes/no queries, select queries that select values of the variables with the matching pattern included, the construction of a graph from these values as a set of RDF statements (subject-predicate-object statements).
An SPARQL query is structured in some parts that resembles SQL (Structured Query Language to query relational databases) with some differences. The basic core of an SPARQL query is a matching pattern facility that uses graph matching pattern functionalities in form of RDF triples. The typical SPARQL query includes the following parts (see Figure 2.6), some of them are optional:

- The PREFIX declarations that allow to abbreviate URIs. For example, the URI <http://www.w3.org/1999/02/22-rdf-syntax-ns#> can be abbreviated as "rdf:"; the URI <http://www.w3.org/2001/XMLSchema#> is abbreviated as "xsd:"; http://www.w3.org/2000/01/rdf-schema# is abbreviated as "rdfs".

- A SELECT clause that includes all the variables in a query to be projected. This clause presents in SPARQL some differences with SQL The SELECT clause in SPARQL can be replaced by the CONSTRUCT clause that returns an RDF graph specified by a graph template, the ASK clause that tests whether or not a query pattern has a solution and the DESCRIBE clause that returns a single result RDF graph containing RDF data about resource.

- A WHERE clause that provides the basic graph pattern to match against the data graph. This graph pattern includes a set of triples that can include variables and operators. These graph patters are the filters for the values to be returned.

- A FROM clause that specifies the graph that is being queried. This clause is optional. If the clause is not specified in the query then, the SPARQL query retrieves information from graphs in an RDF database.

- Query MODIFIERS that order, slice or arrange the returned results obtained from the SPARQL query. For example, the case of ORDER BY that establishes the order of the sequence of results; the LIMIT puts an upper bound on the number of solutions that are returned; the clause OFFSET causes the solutions generated to start after the number of solutions specified in the clause.

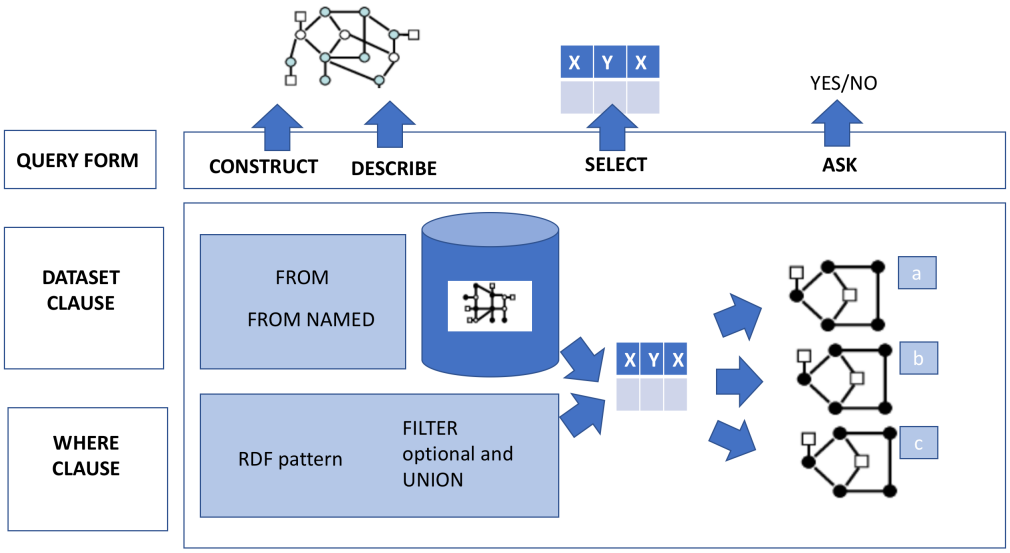

Figure 2.6: The general form of an SPARQL query. The query form can be a SELECT, CONSTRUCT, DESCRIBE OR ASK. The dataset clause specifies the URI or the name of a given graph to be queried. The where clause that provides the RDF pattern (can include the filter optional or union).

The SPARQL syntax is based on IRIs that extends the syntax of URIs to a wider repertory of characters. Therefore, in SPARQL, IRIs are used to design resources. The syntax for literals is a string with either double quotes or single quotes and an optional language tag, IRI or prefixed name. Variables are prefixed with "?" or "$", which are not part or the variable. Blank nodes are designated by a label form such as " _:abc" or "[]". Triple patterns are written as a whitespace-separated list of a subject, predicate and object with their abbreviate forms if they are available.

```
SELECT distinct ?organism ?organame
WHERE {
  ?organism <http://www.w3.org/1999/02/22-rdf-syntax-ns#type>
  <http://khaos.uma.es/pathways/Organism> .
  ?organism <http://khaos.uma.es/pathways/name> ?organame .
}
ORDER BY ?organame
```

Figure 2.7: An example of an SPARQL query that retrieves information from the Kpath endpoint. The SPARQL query code contains a SELECT clause with all variables that were declared, a where clause that includes all the RDF patterns and a query modifier to sort the query results.

Figure 2.7 illustrates an SPARQL query. The first line corresponds with the clause SELECT. In SELECT, the variables ?organism and ?orgname were declared including the modifier DISTINCT

that ensures that duplicate solutions are eliminated. Therefore, the result of the SPARQL query will return a table with two columns with the assigned variables (organism and orgname). The WHERE clause is composed by a pattern of two triples: the first means that the query is looking for all subjects of type "organism". Those predicates can also be typed using rdf:type (abbreviating <http://www.w3.org/1999/02/22-rdf-syntax-ns#> as rdf). The second pattern means to look for organisms with a name. Additional filters (REGEX) can also be added to restrict the results by name but, in this case, it is not specified in the SPARQL query. Finally, the ORDER BY modifier puts the solutions in order according to the organisms' name. The output of this SPARQL query will be a table with two columns with the URIs of the organisms stored in the RDF dataset and their corresponding names. For example: <http://bio2rdf.org/keggtaxon:hsa> as an URI of a given organism and "Homo sapiens" as the name.
SPARQL queries are executed against RDF datasets (see **Definition 3**). Developers and publishers offer an SPARQL service that accepts queries and returns the queries' outputs via HTTP. These results can be returned/rendered in a variety of formats such as XML (that returns the results as tables), a JSON object that is very useful for web applications, RDF (results returned as subject-predicate-objects statements), N-triples (line-based, plain text format that encodes RDF graphs, see [30] for more details of this format), turtle (a concrete syntax for RDF [29]) and HTML.
In this context, with the emerging of the Linked Data technology, the number of services that provide the information in RDF datasets to be queried has grown providing data consumers the opportunity to merge data distributed across the Web. This has caused the creation of a new specification that defines the semantics and syntax of the SERVICE extension to SPARQL version 1.1. Those SPARQL queries that retrieve information from more than one source are called federated queries. These SPARQL queries contain the SERVICE keyword, which instructs a federated query processor to invoke a portion of an SPARQL query against a remote SPARQL endpoint. An example of a federated query processor is ARQ [35], an extension of JENA library that supports federated SPARQL queries.
The following code represents an example of a federated SPARQL query:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX ensembl: <http://rdf.ebi.ac.uk/resource/ensembl/>
PREFIX ensemblterms: <http://rdf.ebi.ac.uk/terms/ensembl/>
PREFIX core: <http://purl.uniprot.org/core/>

SELECT ?uniprot_id ?uniprot_uri ?isoform ?seq {
   ensembl:ENSG00000128573 ensemblterms:DEPENDENT ?uniprot_uri .
   ?uniprot_uri dc:identifier ?uniprot_id .
   SERVICE <http://sparql.uniprot.org/sparql> {
     ?uniprot_uri core:sequence ?isoform .
     ?isoform rdf:value ?seq .
   }
}
```

Figure 2.8: An example of a federated SPARQL query. This query retrieves information from the EBI RDF platform and the Uniprot SPARQL endpoint. This query has the SERVICE clause that invokes a second service to get information about the protein isoforms from the coding gene ensembl:ENSG00000128573.

Figure 2.8 shows an example of an SPARQL query federated query. It retrieves information from more than one source (the EBI RDF platform [36] and the Uniprot SPARQL endpoint). It is composed by several parts: 1) All the prefixes, which are abbreviations that can be used in the query; 2) the SELECT clause that declares all the variables; 3) The pattern of subject-predicate-object statements; 4) The SERVICE keyword that invokes the SPARQL service of Uniprot and

contains a pair of RDF patterns. The query in the example asks the question: "Get the Uniprot ID from the subject ensembl:ENSG00000128573 (a protein coding gene that corresponds with the human CCDS set) and all the isoforms associated with the protein codified by this gene". Thus, the results for proteins at EBI server are complemented with the retrieval of the protein's sequence from Uniprot that are not stored in EBI server.

2.2.3 OWL: The Web Ontology Language

OWL [37] is a standard knowledge representation language for the Semantic Web created and recommended by the W3C consortium. The term knowledge representation refers to the method of modeling the real world by using entities that represent things and relationships between these entities. OWL is a very expressive, flexible knowledge representation language that has had applications in several domains such as health care, traffic and automobiles etc. OWL is based on description logics (DL) [38], which allows that this language can be reasoned. OWL is also based on the Open World Assumption, which means that all is not known to be true is simply unknown and is not necessarily false. The current version of OWL is called OWL 2 [39] that is an extension of OWL [40] and allows more expressiveness than its predecessor. In [41], the authors analyze the identified shortcomings from OWL 1 such as expressivity issues, problems with its syntaxes and definitions of OWL species. Furthermore, the authors present an overview of OWL 2 and how this new version developed by the W3C overcomes the problems presented in OWL 1. In this PhD thesis, we overview the OWL 2 version as we have used it.
The most important element in an OWL 2 ontology is an IRI, which represents a real world entity. As specified in [39], these IRIs represent an ontology and its elements being absolute and not relative. In a given ontology, two IRIs are structurally equivalent if the string construction are identical. IRIs are enclosing by a pair of < (U+3C) and > (U+3E) characters that are not part of an IRI but indicate where it starts or finishes. IRIs can be very long therefore, these are abbreviated as in SPARQL queries with a prefix name pn : followed by an empty string, the : (U+3A) character by associating with a prefix IRI that can be called PI. An IRI called I has a representation as PI, can be abbreviated as $pn : rc$. The prefixes can be specified in the beginning of the OWL document including the authoring of the OWL and be readable by OWL parsers. Classes, properties (relationships between classes), individuals and datatypes are the basic constructs of an OWL ontology. Classes represent concepts of a domain, for example, diseases, pathways, persons etc. In OWL2, there are classes with the IRI *owl:Thing* that represents the set with all the individuals and *owl:Nothing* that corresponds to an empty set. Instances of the classes are called individuals, for example, given a class *owl:Disease* whose symptoms have been modeled in OWL, the individual patient that manifest some of these symptoms can be considered as a individual. Properties represent the relationships between classes. There are two kinds of properties in OWL 2: 1) data properties (OWLDataProperty) that represents the relationship between a class and a datatype that can be a string, integer, boolean. For example, the *owl:Disease* affects to children (with the data property "affectsChildrenhood") is true or false (boolean OWLDataProperty); 2) object properties that relate two classes. For example, the *owl:Disease* affects (*owl:findingSite* as the OWLObjectProperty) "the structure of nervous system" (owl:StructureofNervousSystem). Below there are some constructs of the examples previously described of an OWL ontology based on the domain of the HealthCare:

1. `<owl:Class rdf:about = "Disease"/ >`

2. `<owl:ObjectProperty rdf:about = "findingSite"/ >`

3. `<owl:DatatypeProperty rdf:about = "affectsChildrenhood"/ >`

4. ```
   <rdf:Description rdf : about = "Patient" / >
     <rdf:type rdf : resource = "Disease" / >
   </rdf:Description >
   ```

The first corresponds to the class Disease, the second and the third correspond to the object and datatype properties. The forth corresponds to an instance called “Patient” belonging to the class “Disease". If all these constructs are part of the same ontology, the instance “patient” can have two kinds of relationships “findingSite” and “affectsChildrenhood” with a domain that correspond to the class “Disease” and the ranges for these relationships (ObjectProperty and DataTypeProperty) are a class that defines a part of the human body and a boolean (true or false), respectively.
There are so many domains like the Life Sciences where taxonomy is an essential requirement to model how the data is organized. A taxonomical herarchy can be modeled using classes, subclasses, disjoint axioms, equivalent axioms, object and data properties etc. These constructs can help to model the following statements: 1) The OWL class “Disease” can have have two subclasses called “Mycoses” and “Viral hepatitis”. This means that any instance like “patient” that presents “Mycoses” has also a disease, 2) The classes “Mycoses” and “Viral hepatitis” are not disjoint classes because a given patient can have one of these diseases or both, 3) The class “Mycoses” can have equivalent axioms that are specified with the construct owl:equivalentClass. For example, the class “Mycoses” “hasCausativeAgent” some “Kingdom Fungi” and “hasPathologicalProcess” some “Infectious process”. This axiom can be constructed with two object property relationships “hasCausativeAgent” and “hasPathologicalProcess”. An instance called “Patient”, which has a pathological process and the causative agent belongs to the Kingdom Fungi, has a fungal infection.
In OWL 2 ontologies, there are more constructs that define *more complex classes* that are very useful to the expressively by describing classes that share some properties. Some of these constructs are : 1) *owl:intersectionOf* that is defined as $C \equiv \mathrm{A} \cap \mathrm{B}$ that means individuals in C are also members of A and B, 2) *owl:unionOf* is defined as $C \equiv \mathrm{A} \cup \mathrm{B}$, which means that individuals in C are also members of A, B or both, 3) *owl:complementOf* that is defined as $A^{\complement} = \{x \in U \mid x \notin A\}$ that means if U is the *universe* that contains all the elements and x is not contained in A, then x is contained in $A^{\complement}$.
In addition to these complex class expressions, there are others, which can be used as property restrictions. These property restrictions are used to impose restrictions to a given OWL class. These property restrictions are *owl:allValuesFrom*, *owl:someValuesFrom*, *owl:maxQualifiedCardinality*, *owl:minQualifiedCardinality* and *owl : qualifiedCardinality*. An example of this can be the following construct:

```
<owl:Class rdf:about = "Mycoses"/>
 <rdfs:subClassOf>
 <owl:intersectionOf>
  <owl:Restriction>
        <owl:onProperty>
          <owl:ObjectProperty>
            <rdf:about="hasCausativeAgent"/>
          </owl:onProperty>
        <owl:someValuesFrom rdf:resource="Kingdom Fungi"/>
      </owl:Restriction>
    </owl:Restriction>
  </owl:intersectionOf>
 </rdfs:subClassOf>
```
```

```
</owl>
```

The OWL code above represents the OWL class called "Mycoses", which has as object property "hasCausativeAgent" some (restriction property specified as *owl:someValuesFrom* Kingdom Fungi. The object property in this example has as domain a disease and as range the Kingdom Fungi.
Properties in ontologies play an important role, they describe relationships between OWL classes. Properties are related to one another by using different terms such as *rdfs:subPropertyOf*, *rdfs:equivalentProperty* etc. This means that, semantically, if a property called p_2 is a sub-property of another property called p_1, all the individuals that are held by the property p_2 are also held by the property p_1. In order to construct a very complex relationship, OWL 1 and OWL 2 do not allow constructors such as conjunction, disjunction, etc. However, OWL 2 provides constructs, which are property chains that are very useful in those ontologies where modeling complex properties are necessary. OWL properties have different characteristics such as domains, ranges, reflexive, transitive and asymmetric features, etc.
An example where the OWL properties are specified in an OWL ontology is the following:

```
<owl:Class rdf:about="http://snomed.info/id/279469006">
    <rdfs:subClassOf>
      <owl:Class rdf:about="http://snomed.info/id/279465000"/>
    </rdfs:subClassOf>
    <rdfs:label xml:lang="en">Lumen of penile urethra</rdfs:label>
  </owl:Class>
```

The example above represents that the SNOMED CT class 279465006 from the ontology Dione (OWL 2) corresponds with "Lumen of penile urethra" (body structure) is an *owl:subclassOf* the class 279465000 called "Lumen of male urethra". This means that all the instances included in the class 279465006 belongs to the SNOMED CT class 279465000.
OWL 2 ontologies can take advantage of reasoners to infer knowledge from them. Reasoners are a piece of software, which have been developed in the last few years, some of them are Hermit [42] , Pellet [43], CEL [44], TrOWL [45], Elk [46]. All these reasoners can support visualization tools such as Protégè [47] or NeOn toolkit [48]. These reasoners present several features, which can be enumerated as follows. Ontology satisfiability consists of knowing if the ontology is consistent or not. If a given ontology presents an inconsistency, it refers to the fact the ontology has a contradiction and therefore, two contradictory statements. For example, given an instance a (defined in an ontology O), a is classified by the reasoner in two classes A and B modeled as disjoint classes. Instance checking checks if an instance called a is classified into the class A. Class satisfiability refers if a class A can have instances. Subsumption that checks if two classes called C and D, $C \sqsubseteq D$ (C is a subclass of D). Classification generates all the subclasses'relationships. All these mentioned dimensions are considering in those cases where a comparative analysis with reasoners is performed as in [49]. In this study, the authors provide a complete survey of the existing reasoners and a comparative analysis in terms of classification using large ontologies expressed in DL EL profile such as SNOMED CT [50].
OWL 2 presents different sub-languages in basis of the expressiveness and the scalability requirements. The more expressive an OWL 2 profile is, the more complexity the profile presents. OWL 2 DL profile is the most expressive allowing to retain the completeness, decidability and the availability of practical reasoner algorithms. OWL Lite allows to support users that need a hierarchy classification and simple constraints such as cardinality values between 0 and 1. OWL full is based on the semantics on OWL Lite and OWL DL, with certain compatibility with the RDF schema and no restrictions. OWL Full is undecidable, what means that no reasoner is able to complete the reasoning for it.
OWL 2 EL in which the complexity of inferencing tasks is polynomial. This profile can be very useful for applications that need large number of classes and properties but do not require complex

OWL constructs. OWL 2 QL is designed to support conjunctive query on relational databases. This profile has the worst complexity in term of polynomial time. OWL RL that allows to rule based system to perform reasoning in polynomial time.

2.2.4 The Web of Data

In the subsection above, we have fully described the standard languages to represent and retrieve information (RDF, SPARQL query language and OWL) used in the Linked Data technology. In this subsection, we describe the Web of Data and its properties, the topology of this Web and the Linked Data generated in different domains.
The concept of Web of Data is defined as a global *space* of data generated by a great number of individuals and (private and public) institutions that publish their data. This resulting data space includes all sort of information referring to geography, scientific publications, governments, music, television and radio programs, life sciences etc. The Web of Data presents a set of properties summarized in [51] that can be enumerated here as follows:

- This Web of Data is generic and contains all type of information
- All institutions and individuals are free to publish data in this *space* of linked data
- This global *space* can include disagreement or contradictory information between entities
- All entities are connected through RDF links. All information represented as RDF creates a graph that spans other sources allowing the discovery of new data sources. According to this, applications that consume Linked Data are able to detect new added sources at run-time by following the RDF links that connect entities
- Publishers are not constrained in using a specific vocabulary to represent their data
- Data is self-describing what means that if an application consumes a set of data that follows a unspecified vocabulary, the application must be able to dereference the URIs that identify the terms that represent the concepts
- The application of HTTP as a standard vocabulary taking following the Linked Data principles.

The Web of Data is supported by the Semantic Web community and the Open Linked Data project. This project has as main objective to detect data sources under open source licenses, convert the data to RDF triples and publish the data on the Web. The main principle of this community is that any user can contribute to the increase of the Web of Data. This openesses that characterizes this project can be considered as an important factor in terms of success. This can be empirically confirmed by the the rapid growth that the Web of Data has experienced in the last few years according to the latest statistics provided in [6].
According to [52], the Linking Open Data Cloud available from 2014-08-30 to 2017-01-25 contains the following data sets in the following domains (in parenthesis the percentage of data sets for each domains over the total of data published in the Web of Data): government (23.85%), publications (23.33%), Social Web (15.78%), Life Sciences (11.05%), cross-domain (7.19%), user-generated content (7.36%), geographic (4.21%), media (3.68%), linguistics (3.50%).
Governments and public institutions generate valuable data, which is stored and frequently unused. The governmental data involves information that ranges from economy to citizens' statistics, reports from the public institutions etc. All this information has been increased in the last decade, specially with the tendency of opening the governmental data to the public. Therefore, in this

context, there are some initiatives such as data.gov.uk and the data.gov that interlink their information with other data sets using RDF. The RDF data has been published by these projects following some guidelines about how to publish governmental data proposed by Berners-Lee [53] and Villazón-Terrazas *et al.* [54].
The category of publications holds data sets with information about libraries, scientific publications, citations databases etc. Examples of projects that follow the Linked Data principles are the Swedish Union Catalog called LIBRIS, which started to share linked data in 2008. Other libraries that use Linked Data technology are the British National Bibliography, the Open Library and the German and French National Libraries.
The Linked Data related to the category *user-generated* is generated by portals that collect information from users' communities. Some examples are data about blogposts published as Linked Data by wordpress.com, data from open source softwares like apache.org, workflows in Life Sciences published in myExperiment.org , and reviews generated in goodreads.com and revyu.com.
The category of social network includes RDF data from social users' profiles and the connections amongst these people, including FOAF profiles and the ties generated by StatusNet. The category user-generated is different to the social network category as the former contains information about data published by users in pages that involve a community and the latter about public users'social profiles.
The cross-domain data, which is another category of the Linked Data Cloud, refers to information that is not specific of a given topic. This type of domain for data is very necessary for connecting specific data sets to others avoiding non-interconnected fragments of the Linked Data Cloud in terms of topics. An example of this cross-domain Linked Data is DBpedia [55], which is a project that extracts structured information from Wikipedia [56] making this information accessible on the Web of Data. From the beginning of this project, DBpedia has served as an central *hub* for the emerging of the Linked Data. Furthermore, in the last few decades, the main focus of DBpedia has been extracted information from Wikipedia articles from infoboxes, images, categorization of the information, citations, links etc. Further cross-domain data sets include UMBEL (Upper Mapping and Binding Exchange Layer), a logically organized knowledge graph, linguistic resources such as WordNet or Lexvo and product data.
The data sets in the category *geographics* relate several domains that have a background about geography. For example, GeoNames was the first data set to offer geographical information as Linked Data and to serves as a *hub* in this topic. Another significant data set, which is part of the geographic Linked Data is LinkedGeoData. This RDF data set collects information from the OpenStreetMap (a source of spatial data) and makes it available by REST services. DBpedia also provides geospatial RDF information and is interlinked with GeoNames and LinkedGeoData, contributing to the increasing of the Linked Data Cloud [57].
The category *media* contains data sets that provide information about music, films, TV and radio. One of the most prominent data sets within this category is the British Broadcasting Corporation (BBC) that publishes large amounts of content on the Web such as texts, videos and audio. However, most of this information is accessible through specific HTML microsites, excluding a wider data integration with the rest of the BBC information. This context involves problems related to the cross-domain linking and the disambiguation between vocabularies. Therefore, in order to address these problems, the authors in [58] propose to use the DBpedia, which provides a standard vocabulary to be interlinked with the BBC data following the Linked Data principles. The resulting work led to a categorization system called CIS to interlink data items and a set of services that use underlying Linked technology. This project can benefit users through topic pages and navigation badges. Likewise, another important data set within this category that follows the same approach is The New York Times that publishes pieces of information as RDF by interlinking data with existing data sets such as DBpedia, GeoNames etc.
Life Sciences has become a very prominent branch that has been generating amounts of information

as technologies have advanced in the post-genomic era. All this information contained in isolated databases has been turned to RDF according to the Linked Data principles, contributing to the Web of Data. Examples of prominent projects in the Life Sciences Linked Data are Bio2RDF [25] and BioPortal [59]. Bio2RDF is a system that helps to solve the problem of knowledge integration in the Life Sciences domain [25]. The Bio2RDF project can be seen as a mush-up because it combines information from 35 different Life Sciences data sets. The number of RDF triples that has been generated by Bio2RDF corresponds to 11 millions. Another project is BioPortal, an open repository of biomedical ontologies that stores more than 700 developed ontologies in different formats such as OWL, OBO, RRF. In [60], the authors published an RDF version of BioPortal that has initially more than 190 millions of RDF triples. Both projects have contributed to enrich the web *space* in the category Life Sciences. All these contributions have performed by a users' community that comprises researchers, organizations etc. allowing the evolving of the *space* of data to the current status.

2.2.5 Linked Data Applications

After the emergence of the Linked Data technology, the so-called Linked Data applications have had an important role. According to [61], there are two ways to define the term of *Linked Data Applications*; the first refers to applications of Linked Data in specific domains such as Life Sciences, *media* etc. and the second corresponds to applications that are implemented on the top layer of the Linked Data. Both types of interpretations about Linked Data applications do not exclude each other and must be complementary. Therefore, we can define the Linked Data applications as those that are designed to consume and manipulate Liked Data. These Linked Data applications can be divided in two groups: 1) generic applications and 2) domain-specific applications. The first group of applications can process information from any domain and the second refers to applications that consume Linked Data from a specific domain such as Life Sciences, Health Care, Cross-domain, Libraries etc. Amongst the generic Linked Data applications, there are also two categories: the Linked Data browsers and the Linked Data engines.
The Linked Data browsers were one of the first implemented systems that took advantage of the Linked Data technology. Similarly to the traditional ones, these browsers navigate through the Web by using RDF links and represent the RDF resources and their properties. For example, Haystack is an application implemented in 2004 and aggregates RDFs from multiple locations allowing users to navigate through RDFs. Haystack displays this information by cascading sytlesheets, which are described in RDF [62]. This browser also provides a rich model of collections to gather and structure information. Disco Hyperdata [63] allows users to render RDF information as HTML pages. Noadster [64] provides a hypermedia document interface to information encoded in Semantic Web standards. This application provides information request interface that returns a list of starting point to navigate through. Piggy Bank [65] is a web-browser extension that restructures information in HTML pages into RDF to generate information in Web Semantic format to be consumed by applications. LESS [66] is an end-to-end approach for the syndication and use of Linked Data based on the definition of templates for linked data resources and SPARQL query results. These syndication templates are edited, published and shared by using a collaborative Web platform. Tabulator [67] is an RDF browser designed for end-users to interact and navigate through RDF resources and for developers of RDF content as incentive for them by posting and refining their information in RDF and show how this new RDF content interacts with other information. LENA [68] provides different views of data, following user's criteria that are expressed as SPARQL queries. Visor [69] is a multi-pivot based browser, which can be configured in any SPARQL endpoint. This tool allows users to explore specific classes and properties from collections' RDF schema and create spreadsheets with the selected information. Other faceted browsers are Humboldt [70], facet [71] and gFacet [72].

Explorator [73] is an open-source exploratory search tool for RDF information, implemented in a direct manipulation interface metaphor. It implements a set of custom operations and some examples of SPARQL queries. Information Workbench [74] is another exploratory tool for RDF that offers several back-end and front-end tools. Marbles [75] is a project that uses Fresnel vocabulary to render RDF information as HTML pages.
As the amount of RDF information has increased with the emerging of the Linked Data technology, searching data and providing useful information in any user's request is a challenge. A number of search engines, categorized as generic Linked Data applications, crawl the Web of Data and provides aggregated information. Examples of these efforts are Swoogle [76] that provides search over RDF documents by a inverted keyword index and a relational database. Swoogle has also some metrics to calculate the popularity of classes and relationships. Watson [77] another semantic browser, which is very similar to Swoogle and provides a keyword-based search to find documents and an API service. Sindice [78] offers a lookup service based on Lucene and MapReduce. The Falcons browser provides an entity-centric searching for concepts over RDF data. SWSE [79] is another search engine that operates directly over RDF data and consists of a crawling, data enhancing, indexing and a user-interface to display results. With all these available semantics browsers, other studies have focused on applying new approaches to ranking the retrieved information from the search as is proposed in [80].
As we have pointed above, there is a second group of Linked Data applications that correspond to applications that cover the needs of specific users' communities. For example, in the category of *media* in the Linked Data Cloud, one of the most popular is the BBC Linked Data platform. The BBC is one of the first organizations in using Linked Data by supporting events related to sports, education, music etc. by means of a set of ontologies such as the BBC ontology, the Food ontology etc. In the category of cross-domain in the Web of data, an interesting project is IBM Watson [81], which is an AI platform able to understand complex questions and answer them. The platform incorporates knowledge databases that follow the standards of the Linked Data technology. Another example in the category of the social network is the Facebook's graph API, a project that presents a consistent view of the social Facebook's graph where users' profiles are presented as nodes and their relationships as arcs. This output information formatted as JSON was converted to RDF/Turtle, which is semantically richer than JSON format, making it accessible as ontologies are based on the RDF standard [82]. Finally, Google [83] also uses Linked Data formats to display Rich Snippets for Web pages, improving how the results are displayed, having an important impact in terms of economy.
In this subsection, we have defined the concept of Linked Data *applications*. We have also classified them into two groups called generic or specific-domain Linked Data applications. Some examples of these Linked Data applications have been provided for each group. In the next subsection, we describe how Linked Data in the domain of Life Sciences has evolved in the last decade and the Linked Data applications implemented over this period, including a project that was chosen as research work to support this PhD dissertation.

2.2.6 Linked Data and Life Sciences

The Life Sciences field has entered in a new era of *Big Data* with the breakthroughs in sciences and technology. This fact has incremented the interest of the scientific community in the Linked Data technologies. These technologies allow users to apply these standards to the generated biological data, integrate them and publish them as Linked Data.
Some efforts in the field of Life Sciences have been materialized in projects such as Bio2RDF. In this work, Belleau *et al.* proposed a mash-up system that tries to solve the problem of the data integration in Life Sciences [25]. Bio2RDF integrates data from most popular databases such as Kegg, UniProt and Reactome. All the integrated information is displayed in an graphical

interface where end-users can navigate through keyword-based searches and RDF links. Another prominent project is LODD (Linking Open Drug Data) where Samwald *et al.* surveyed all the publicly available data about drugs from repositories such as ClinicialTrials.gov and DrugBank and converted them into RDF [84]. Linked Life Data (LLD) [85] is one of the first projects based on the integration of biological data that used Linked Data. This platform currently stores 10 billions of biomedical RDF statements in the field of Life Sciences and Health Care. The platform interconnects up to 25 data sources such as ChEBI, DrugBank, PDB, Pfam, PubMed and Uniprot. There are many data providers that have published their biological data as RDF. For example, BioGateway provides a resource that integrates data from OBO foundry candidate ontologies, GO annotation files, Swiss-prot and NCBI taxonomy. This data has been converted to RDF and is accessible through an SPARQL endpoint where the information can be retrieved [86]. LinkDB is another relational database, which integrates data about compounds, genes etc. that was converted to RDF and can be accessed by an SPARQL endpoint [87] or a client. The PDBj (Protein Data Bank Japan) is a database of proteins' atomic coordinates that provides all its information in RDF [88].
In 2014, the EBI that is the largest bioinformatics resource in Europe, coordinated the effort to bring together RDF resources from multiple data sources and services [36]. The aim of such a platform is to offer users the ability to ask questions using multiple connected resources that share common identifiers and have a common format (RDF) and a query interface (SPARQL). After the launch of the EBI RDF platform, in the last few years, other projects have contributed to the Linked Data initiative. For example, in 2014, Biomodels Linked dataset was published as a dereferencable interlinked dataset that exposes the metabolic and biosignaling models' content [89]. In 2015, GlycoRDF was the first repository that exposes information about glycomics in RDF [90]. At the same year, the National Library of Medicine investigated the potential of publishing MeSH (Medical Subjects Headings) in RDF [91]. In 2016, PubChemRDF was created as an RDF repository that integrates data from three interlinked databases such as Substance, Compound and BioAssay. In 2017, DisGeNET was created as a platform that offers RDF information about human-diseased and variation from different data sources such as GWAS catalogues, expert curated repositories, animal models and scientific literature. All this information is accessible through a Web interface, a Cytoscape application, an RDF SPARQL endpoint, scripts in several programming languages and an R package [92].
In this subsection, we have described the contributions in Linked Data to the category of Life Sciences. All these mentioned projects seem very promising and are the best illustrations about how Life Sciences has adopted the Linked Data technology. In the next subsection, we stress the importance of ontologies (OWL) as an extension of RDF in the fields of Life Sciences and Health and all the studies performed to formalize standardized diseases'classifications (such as ICD-10-CM) as OWL representations.

2.2.7 Ontologies and Biomedicine: the Case of ICD-10-CM

Over the last decade, ontologies and terminologies in the domain of the Life Sciences, specially in the field of biomedicine, have played an important role through a variety of applications such as data management and interoperability, data integration and reasoning and decision support according to Bondenreider's analysis [93]. These applications have conducted to ontologies and terminologies such as SNOMED CT [50, 94], FMA [95], Gene Ontology [96], UMLS [97] and the Medical Subject Headings (MESH)[98].
In the role of data management, biomedical ontologies and terminologies serve as standard vocabulary sources. For example, SNOMED CT has been used to annotate electronic health records [50] and the ICD-10 classification has been used for indexing with automatic techniques [99, 100, 101]. Certainly, for clinical decision support, biomedical ontologies play an important role as they are

considered to be a source of computable knowledge and a standard vocabulary that can be applied to guide medical experts in decision making. An example of the use of an ontology for clinical decision support is the study carried out by [102] in which a selective lung cancer treatment was evaluated through reasoning with the LUCADA ontology, which was developed by the authors and mapped to SNOMED CT. BioPortal is also a very important repository that include all the ontologies in Life Sciences previously published [59].

ICD-10-CM is a standard diagnostic tool for health management, epidemiology and clinical purposes that has been widely used for annotations in database entries from Online Mendelian Inheritance in Man (OMIM) [103] and Orphanet [104]. ICD-10 comprises Chapters I to XXII, which cover diseases, a variety of signs and symptoms, abnormal findings, complaints, social circumstances and external causes of injuries and diseases. ICD-10 contains definitions, inclusions and exclusions for each disease's category.

The ICD-10-CM corresponds to the tenth version, clinical modifications. This medical classification standard, maintained and published by the World Health Organization (WHO) is used to classify diseases and health problems that have been recorded on death certificates and also in other records. The accuracy of this classification is a very important issue because it is used, for example, to set capitation rates and allocate resources to medical centers or to determine the case fatality and morbidity rates by medical health and health services researchers. The development of tools that support a semi-automatic classification would clearly improve the accuracy of the process. Current classification processes are usually done manually.

According to the reviewed literature, there have been several attempts to model ICD-10 as an OWL ontology. For example, the work developed by [105] was the first attempt to model the ICD-9 ontology, an older version of the current ICD-10. In [106], the authors proposed the first formal representation of the ICD-10 based on three logical layers of the GALEN Core Reference Model (CRM) terminology system. They used a description logic-like language called GRAIL which allows classes to be inferred with the semantics of role propagation and links a more detailed description of a diagnosis to a more abstract class. In 2008, the same authors proposed a DOLCE-based formal representation [107]. DOLCE is a descriptive upper-level ontology designed for ontology cleaning and interoperability. In this formal representation of the ICD-10, anatomical entities were taken from the Foundational Model of Anatomy (FMA), morphological abnormalities and procedures were taken from SNOMED CT, the organisms used were from the biological taxonomy and the chemical objects were taken from the International Union of Pure and Applied Chemistry nomenclature (IUPAC). Finally, the last approach to represent ICD-10 in OWL was developed by [108] where an ontology with an OWL-full expressiveness was proposed with the ICD-10 exclusions modeled using disjoint OWL classes. In this PhD dissertation, we have proposed a new approach based modeling ICD-10-CM with axioms that handle definitions, inclusions and exclusions by the ICD-10-CM/SNOMED CT mappings [4].

Chapter 3

Published Work

We have published several research studies based on applications of semantics in Life Sciences. Specifically, three articles were published in journals indexed in the Journal of Citation Report (JCR) from the Institute of Scientific Information. Furthermore, four additional articles were presented in congresses, being two of them in international conferences.

3.1 List with Research Contributions

The publications that emerged from this thesis can be organized as follows according to their type of publication:

Articles published in journals indexed in JCR:

- M. J. García Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Sharing and executing linked data queries in a collaborative environment". *Bioinformatics* 29.13 (2013), pp. 1663–1670. DOI: 10.1093/bioinformatics/btt192
  Impact Factor (2013): 4,621. Q1 (2/57) in the category of Mathematical and Computational Biology.

- I. Navas-Delgado, M. J. Garcia-Godoy, E. Lopez-Camacho, M. Rybinski, A. Reyes-Palomares, M. A. Medina, and J. F. Aldana-Montes. "kpath: integration of metabolic pathway linked data". *Database* 2015.0 (2015), bav053–bav053. DOI: 10.1093/database/bav053
  Impact Factor (2015): 1,548. Q1 (8/56) in the category of Mathematical and Computational Biology.

- M. del Mar Roldán-García, M. J. García-Godoy, and J. F. Aldana-Montes. "Dione: An OWL representation of ICD-10-CM for classifying patients' diseases". *Journal of Biomedical Semantics* 7.1 (2016). DOI: 10.1186/s13326-016-0105-x
  Impact Factor (2016): 1,845. Q2 (18/57) in the category of Mathematical and Computational Biology.

Articles published in international congresses:

- M. J. García-Godoy, I. Navas-Delgado, and J. Aldana-Montes. "Bioqueries: A Social Community Sharing Experiences While Querying Biological Linked Data". *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*. SWAT4LS '11. London, United Kingdom: ACM, 2012, pp. 24–31. ISBN: 978-1-4503-1076-5. DOI: 10.1145/2166896.2166906. URL: http://doi.acm.org/10.1145/2166896.2166906

- M. J. García-Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Reconstructing Hidden Semantic Data Models by Querying SPARQL Endpoints". *Database and Expert Systems Applications: 27th International Conference, DEXA 2016, Porto, Portugal, September 5-8, 2016, Proceedings, Part I.* ed. by S. Hartmann and H. Ma. Cham: Springer International Publishing, 2016, pp. 405–415. ISBN: 978-3-319-44403-1. DOI: 10.1007/978-3-319-44403-1_25

Articles published in national conferences:

- M. J. García Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Bioqueries: a Social Community for SPARQL queries in Life Sciencess". *Actas de las XIX Jornadas de Ingeniería del Software y Bases de Datos. JISBD September 16-19.* 2014

- M. J. García Godoy, E. López-Camacho, M. d. M. Roldán-García, and J. F. Aldana-Montes. "Enriquecimiento Automático de Ontologías Biomédicas mediante el uso de Mappings". *Actas de las XXIII Jornadas de Ingeniería del Software y Bases de Datos. JISBD September 17-19.* 2018

3.2 Summary of the articles that support the thesis

This section summarizes the articles that support this thesis. All these papers present some applications of semantics in the life sciences domain. The first and the second article presents Bioqueries, a tool implemented to stimulate the consumption of biological linked databases and to ease the use of the SPARQL language to query their data. The third article introduced a tool to automatically explore the underlying structure of a linked data repository, making easier the design of SPARQL queries. In the fourth article, we present DIONE: the first OWL representation of ICD-10-CM, which can be considered as a first step towards the automatic classification of patients' diseases by using the Linked Data technology.

3.2.1 Bioqueries: a social community sharing experiences while querying biological linked data

Reference: [1] M. J. García-Godoy, I. Navas-Delgado, and J. Aldana-Montes. "Bioqueries: A Social Community Sharing Experiences While Querying Biological Linked Data". *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences.* SWAT4LS '11. London, United Kingdom: ACM, 2012, pp. 24–31. ISBN: 978-1-4503-1076-5. DOI: 10.1145/2166896.2166906. URL: http://doi.acm.org/10.1145/2166896.2166906

This work [1] presented Bioqueries[1] in the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences (SWAT4LS 2011), celebrated in London in December of 2011. Bioqueries is a tool presented as a social community where users are able to design, document and execute queries from different biological SPARQL Endpoints. This tool was designed taking into account two different profiles of users: a bioinformatic profile which would be able to introduce new SPARQL queries and a biological profile user who can made use of these already created queries and consult the existing biological data.
The tool is presented as a web application in the common format of a wiki. When users are registered into the application, they gain the ability to create their own SPARQL queries that will be executed to a collection of existing external public linked data endpoints, that are already

[1] http://bioqueries.uma.es

included in the platform by the administrators. Each created SPARQL query can be documented and has to be specified as an human-readable parametrized statement. This way, when users desire to execute the query, they will encounter an easy form to fill form instead of a SPARQL query code when they want to consult a external database. The result of each query is presented in a table and can be exported in several common linked data formatted files. Each user can edit and execute their created queries, and they have the capability to make them public to other users in order to be executed and shared with the rest of the community.
To encourage new users to start using our platform, Bioqueries was populated with a set of 116 public predesigned queries. All the queries were categorized according to the biological nature of the queried data. Users can consult and look for queries in the platform browsing by their categories, endpoints or using a common text search.
The Bioqueries web application was implemented using Drupal. For executing the SPARQL queries, we created our own REST service using Java and the Jena libraries for manipulating linked data and querying SPARQL endpoints.

3.2.2 Sharing and executing linked data queries in a collaborative environment

Reference: [2] M. J. García Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Sharing and executing linked data queries in a collaborative environment". *Bioinformatics* 29.13 (2013), pp. 1663–1670. DOI: 10.1093/bioinformatics/btt192

In [2], we presented an improved version of Bioqueries taking into account the suggestions received in the SWAT4LS conference. We also made an study of the usability and usage that our platform had taken in order to measure if our main objective was reached: the creation of an active community of biologists built around the use of Biological Linked Data.
The main new feature that we presented in these articles was the possibility of create federated queries. Federated queries, unlike normal queries, can extract data from more than one data source at the same time. These types of queries make one or several SPARQL protocol calls within a query, splitting it into sub-queries that are sent to each partipating endpoint. The different results are combined when all the sub-queries are executed and finally they are shown to the user.
Additionally, the Relfinder software [112] was included in Bioqueries to ease the process of constructing a SPARQL query. Relfinder is used to visualize relationships between two or more RDF nodes from a selected repository and find new unknown relationships. These relations are presented in an interactive graph viewer.
Sometimes, the availability of the included endpoints is temporally limited because of server maintenance, server failure, network problems and so forth. This situation is beyond the control of administrators of Bioqueries; therefore, a panel with the endpoint status information was included. Additionally, before an user tries to execute a query, Bioqueries tests automatically the corresponding endpoint status and disallows queries execution while the endpoint is unavailable.
We also carried out a study of the usage of the platform since it was made public one year and a half before. The number of queries reached to 215 (being around 5.6% made by external users) and the registered users were 230. The queries in the platform were viewed >13368 times (excluding those made by their own authors). A usability assesment was also performed using a variety of questionaries and then calculating the SUS score [113]. This score measures the effectiveness (the ability of users to complete the tasks), efficiency (the level of resources consumed to carry out the task) and the users' satisfaction (a user's reaction derived from the system use). The questionary was solved by a collection of 12 participants, being 5 of them with biological profiles and 7 of them with bioinformatic profiles. The obtained SUS score was 78.3 and 79.6 respectively.

3.2.3 Re-constructing Hidden Semantic Data Models by Querying SPARQL Endpoints

Reference: [3] M. J. García-Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Re-constructing Hidden Semantic Data Models by Querying SPARQL Endpoints". *Database and Expert Systems Applications: 27th International Conference, DEXA 2016, Porto, Portugal, September 5-8, 2016, Proceedings, Part I.* ed. by S. Hartmann and H. Ma. Cham: Springer International Publishing, 2016, pp. 405–415. ISBN: 978-3-319-44403-1. DOI: 10.1007/978-3-319-44403-1_25

This article was presented in the 27th International Conference on Database and Expert Systems Applications (DEXA 2016) celebrated in Porto, Portugal in September of 2016. It was derived from the difficulties we encountered when designing SPARQL queries. Ideally, the underlying semantic model of the data should be documented using the standard RDF Schema vocabulary VoID [2] or through more simple approach like an HTML documentation. However, that is not always the case. Because of that, we developed a tool that automatically reconstruct the semantic data model behind an RDF database.
The approach this tool uses is based on a set of SPARQL queries that are executed to infer the implicit RDF structure of a given endpoint. In a first step, some queries are executed to retrieve the set of classes and properties contained. Then, two queries will be executed for each property in order to extract their domain and range. According to the range of the property, it is possible to divide the obtained list of properties into two subsets: object properties (when its range is composed by classes) and data properties (when its range is composed by datatypes). Finally, the result of these queries are combined to produce a single OWL ontology that represents the internal semantic model of the data.
This tool was developed using Java, and its code is publicly available in GitHub [114]. We also implemented a web application [115] that provides all the functionalities of the tool, so users do not have to compile and execute the code. The obtained results can be viewed using this web or can be exported and downloaded in an OWL file.
This tool was tested using four use cases: two SPARQL endpoints developed by us (Kpath [109] and ReprOlive [16]), a well-known SPARQL endpoint in the Life Sciences domain (Biomodels [17]) and a well-known SPARQL endpoint in the Linked Data community (LinkedGeoData [15]). These four examples were selected according to its data size and complexity in order to measure the quality of the obtained ontology. The created ontologies in these examples can also be downloaded from the webpage of the tool.

3.2.4 Dione: An OWL representation of ICD-10-CM for classifying patients' diseases

Reference: [4] M. del Mar Roldán-García, M. J. García-Godoy, and J. F. Aldana-Montes. "Dione: An OWL representation of ICD-10-CM for classifying patients' diseases". *Journal of Biomedical Semantics* 7.1 (2016). DOI: 10.1186/s13326-016-0105-x

This article was published in the Journal of Biomedical Semantics. In this article, we have focused on the importance to formalize biomedical terminologies such as ICD-10-CM using the OWL standard. According to the reviewed literature, there have been some attempts that have failed to model the ICD-10-CM as an OWL representation. In one of the latest studies performed by Möller *et al.* [108], the authors presented a model that captures the hierarchical information of ICD-10

[2] https://www.w3.org/TR/void/

and handles the ICD-10-CM inclusions and exclusions with an OWL full component. Although the authors suggest that the part of the ontology is expressed by using the OWL-DL profile and an OWL full component for those properties that exceed the OWL-DL expressivity, the proposed model has not been tested in an OWL reasoner. This can be explained as the model has an OWL full component that means that reasoners cannot complete the reasoning for it.

Therefore, according to the reviewed literature and the lack of OWL models for ICD-10, we have proposed Dione, which is the first OWL that is logically consistent, whose axioms that define the ICD-10-CM exclusions and inclusions are based on the SNOMED CT/ICD-10-CM mappings. These mappings were extracted from the official SNOMED CT/ICD-10-CM mappings from UMLS. The axioms extracted from these mappings complete Dione in a 93%. That means that the 93% of the diseases'classes has axioms. Therefore, in order to provide a more complete version of ICD-10-CM, we used other mappings from BioPortal that were validated in a semi-automatic way. These new BioPortal mappings allowed us to generate a version of Dione where the mappings completed it in a 93,3%.

Dione currently contains 391,669 classes, 391,720 entity annotation axioms and 11,795 owl:equivalentClass axioms. The resulting OWL representation has been classified and its consistency tested with the ELK reasoner. Therefore, in order to put in practice Dione, we have also taken three clinical records from the Virgen de la Victoria Hospital (Málaga, Spain), which have been manually annotated used SNOMED CT. These annotation have been included as instance to be classified by th Elk reasoner. The classified instances show that Dione could be a promising ICD-10-CM OWL representation to support the classification of patients' disease

3.3 Summary of other publications related to this thesis

This section briefly comments the other three articles that do not support this thesis but are related to its topic. One of them was published in the Database journal and the other two were published in different years of the JISBD national conference.

In [109], we presented kpath, a database that integrates information related to metabolic pathways. This database uses the Linked Data approach to solve the heterogeneity problems of the available biological databases and to enable the integration of additional data sources with a controlled cost. We also presented the kpath browser[3], a client interface that provides three different tools (Pathway Graphical Viewer, Pathway Graphical Editor and Relationship Search Tool). Their combination provides users with not only a browsing interface, but also some analytical features to discover new knowledge from the integration of public data. We also provided our data as an Open SPARQL endpoint[4], allowing external parties to develop different user interfaces on top of this database.

In [111], we presented an authomatic methodology that allows the Dione ontology to be populated with axioms created from established mappings between ICD-10-CM and another biomedical ontology that are provided by BioPortal. In this paper, we used as an example the mappings between Dione y ORDO. ORDO is an ontology that includes rare diseases, genes and other additional features. Using this authomatic process is possible to maintain the ontologies updated with the lattest mappings included in Bioportal and allow more complex reasonings that would be impossible otherwise.

Finally, in [110], we used the opportunity in the XIX Jornadas de Ingeniería del Software y Bases de Datos (JISBD 2014) to spread Bioqueries among the database community.

[3] http://browser.kpath.khaos.uma.es/
[4] http://sparql.kpath.khaos.uma.es/

3.4 Copies of the articles that support the thesis

In this section, the list of published papers that support this thesis is enumerated as follows:

M. J. García-Godoy, I. Navas-Delgado, and J. Aldana-Montes. "Bioqueries: A Social Community Sharing Experiences While Querying Biological Linked Data". *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*. SWAT4LS '11. London, United Kingdom: ACM, 2012, pp. 24–31. ISBN: 978-1-4503-1076-5. DOI: 10.1145/2166896.2166906. URL: http://doi.acm.org/10.1145/2166896.2166906

Motivation: Life Sciences have emerged as a key domain in the Linked Data community because of the diversity of data semantics and formats available through a great variety of databases and Web technologies. Unfortunately, bioinformaticians are not exploiting the full potential of this technology and experts in Life Sciences have real problems to discover,understand and devise how to take advantage of these interlinked data.
Results: In this context, we have implemented Bioqueries, a wiki-based portal that is aimed at community building around biological Linked Data. This space offers a collaborative platform in which users can create, modify, execute and share biological SPARQL queries.
Availability and implementation: https://www.bioqueries.uma.es

M. J. García Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Sharing and executing linked data queries in a collaborative environment". *Bioinformatics* 29.13 (2013), pp. 1663–1670. DOI: 10.1093/bioinformatics/btt192

Motivation: Life Sciences have emerged as a key domain in the Linked Data community because of the diversity of data semantics and formats available through a great variety of databases and Web technologies. Unfortunately, bioinformaticians are not exploiting the full potential of this technology and experts in Life Sciences have real problems to discover,understand and devise how to take advantage of these interlinked data.
Results: In this article, we present Bioqueries, a wiki-based portal that is aimed at community building around biological Linked Data. This tool has been designed to aid bioinformaticians in developing SPARQL queries to access biological databases exposed as Linked Data, and also to help biologists gain a deeper insight into the potential use of this technology. This public space offers several services and a collaborative infrastructure to stimulate the consumption of biological Linked Data and, therefore, contribute to implementing the benefits of the web of data in this domain. Bioqueries currently contains 215 query entries grouped by database and theme, 230 registered users and 44 end points that contain biological Resource Description Framework information.
Availability and implementation: The Bioqueries portal is freely accessible at http://bioqueries.uma.es

M. J. García-Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Reconstructing Hidden Semantic Data Models by Querying SPARQL Endpoints". *Database and Expert Systems Applications: 27th International Conference, DEXA 2016, Porto, Portugal, September 5-8, 2016, Proceedings, Part I.* ed. by S. Hartmann and H. Ma. Cham: Springer International Publishing, 2016, pp. 405–415. ISBN: 978-3-319-44403-1. DOI: 10.1007/978-3-319-44403-1_25

Motivation: Linked Open Data community is constantly producing new repositories that store information from different domains. The data included in these repositories follow the rules proposed by the W3C community, based on standards such as Resource Description Framework (RDF) and the SPARQL query language. The main advantage of this approach is the possibility of external developers accessing the data from their applications. This advantage is also one of the main challenges of this new technology due to the cost of exploring how the data is structured in a given repository in order to construct SPARQL queries to retrieve useful information. According to the reviewed literature, there are no applications to reconstruct the underlying semantic data models from an SPARQL endpoint.
Results: In this paper, we present an application for the reconstruction of the data model as an OWL (Ontology Web Language) ontology. This application, available as Open Source at `http://github.com/estebanpua/ontology-endpoint-extraction` uses a set of SPARQL queries to discover the classes and the (object and data) properties for a given RDF database. A web application interface has also been implemented for users to browse through classes, properties of the ontology generated from the data structure `http://khaos.uma.es/oee`. The ontologies generated by this application can help users to understand how the information is semantically organized, making easier the design of SPARQL queries.
Availability and implementation: `http://github.com/estebanpua/ontology-endpoint-extraction` and `http://khaos.uma.es/oee`

M. del Mar Roldán-García, M. J. García-Godoy, and J. F. Aldana-Montes. "Dione: An OWL representation of ICD-10-CM for classifying patients' diseases". *Journal of Biomedical Semantics* 7.1 (2016). DOI: 10.1186/s13326-016-0105-x

Motivation:
Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) has been designed as standard clinical terminology for annotating Electronic Health Records (EHRs). EHRs textual information is used to classify patients' diseases into an International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) category (usually by an expert). Improving the accuracy of classification is the main purpose of using ontologies and OWL representations at the core of classification systems. In the last few years some ontologies and OWL representations for representing ICD-10-CM categories have been developed. However, they were not designed to be the basis for an automatic classification tool nor do they model ICD-10-CM inclusion terms as Web Ontology Language (OWL) axioms, which enables automatic classification. In this context we have developed Dione, an OWL representation of ICD-10-CM.
Results:
Dione is the first OWL representation of ICD-10-CM, which is logically consistent, whose axioms define the ICD-10-CM inclusion terms by means of a methodology based on SNOMED CT/ICD-10-CM mappings. The ICD-10-CM exclusions are handled with these mappings. Dione currently contains 391,669 classes, 391,720 entity annotation axioms and 11,795 owl:equivalentClass axioms which have been constructed using 104,646 relationships extracted from the SNOMED CT/ICD-10-CM and BioPortal mappings included in Dione using the owl:intersectionOf and the owl:someValuesFrom statements. The resulting OWL representation has been classified and its consistency tested with the ELK reasoner. We have also taken three clinical records from the Virgen de la Victoria Hospital (Málaga, Spain) which have been manually annotated using SNOMED CT. These annotations have been included as instances to be classified by the reasoner. The classified instances show that Dione could be a promising ICD-10-CM OWL representation to support the classification of patients' diseases.
Availability and implementation: Dione is available at http://www.khaos.uma.es/dione

Chapter 4

Conclusions and Future Work

This chapter exposes the final ideas of this dissertation, the conclusions obtained in all the past experiments and the future lines of work that we plan to explore from the latter works.
The amount of data on the Web has enormously increased over the last two decades, specifically, in the domain of the Life Sciences. This amount of information causes the necessity of including semantics to be processed by machines. The technology of the Linked Data emerged in 2008 and can be defined as a set of recommended practices for sharing, exposing and connecting pieces of information, data and knowledge by using URIS, RDF and OWL standards. This technology, supported by the W3C community, has been applied to different categories of data such as Government, Publications, Social web, Cross-domain, Geographic, Media, User-generated and Life Sciences.
Life Sciences was one of the earliest adopters of the Linked Data technology. This adoption has been materialized due to the great amount of Life Sciences data that has been generated with all the breakthroughs in science and technology. However, the adoption of this technology has presented some problems such as: 1) the availability of the RDF data sets 2) the heterogeneity of semantics and 3) the steep learning curve of the Life Sciences researchers.
Therefore, according to the problems previously specified, this thesis proposes several solutions that combine the study of each problem with the software development. Bioqueries is the first contribution to this thesis, providing a collaborative environment for the community of the Life Sciences. This environment has a relative success that is shown by the high number of users. However, one of the shortcomings is related to the low activity of the users' community. This is explained by the low rate of query publication (only 5,6% of queries). However, the number of times queries are executed is high and this can be explained with the effort in the design of a significant number of queries and the development of tools such as the detection of the availability of RDF repositories. Given the problem of the low activity related the query publication in Bioqueries, we have attempted to overcome the problem of the extraction of the hidden semantic model from RDF repositories. In an ideal scenario, the model should be documented but in practice, the RDF developers do not provide it. To overcome this problem, we have designed and implemented an algorithm that allows to reconstructing automatically the semantic model behind an RDF database [3]. The resulting tool is public and available to the scientific community at [114] as well as a Web service where the algorithm can be executed and the results displayed [115]. The algorithm was tested by carrying out some experiments that involved RDF repositories such as LinkedGeoData, Biomodels, ReprOlive and kpath. However, despite the algorithm retrieved the underlying semantic model from the repositories, the retrieved model is partial. An example of the problem is that the subsumption relationship between classes is not obtained if that relationship has not been previously specified in the repository. Therefore, we are currently working on a second

version of the approach based on obtaining these relationships through two strategies: 1) applying an alignment of the extracted ontologies with existing ones and then, completing the relationships that have not been retrieved and 2) extracting of the instance set for each discovered class in the endpoint and calculating if any of them have a subsumption relationship. The process can be manually curated to detect false positive. Furthermore, we are planning as s post-doc research line the integration of this technique with Bioqueries, trying to increase the ratio of users'activity. Furthermore, a feasible solution for the problem Bioqueries presents is the incorporations of new features that help end-users to create semi-automatically SPARQL queries and the inclusion of more complex queries consolidating Bioqueries as universal repository of SPARQL queries in the Life Sciences. Ontologies are part of the standards that Linked Data technology encourages to apply to the generated data. These standards have been widely used in domains such as Health Care and Life Sciences. Ontologies, according to their existing profiles, allow to be reasoned by OWL reasoners. This feature has provided ontologies an important role in biomedicine as classification systems for categories of diseases by using standard biomedical terminologies. An important challenge has been the modeling the ICD-10 as an OWL representation. According to the literature, there have been so many attempts to model the ICD-10 but all of them have failed to implement an OWL ontology that is logically consistent. This has motivated us to create Dione as the first OWL representation to model the ICD-10-CM, which is the latest version of the ICD. For performing this, we have implemented an automatic process to construct the hierarchy tree with ICD-10-CM disease classes and their axioms with *owl:equivalentClass* axiom. The result from this process was an OWL model that can be used by a reasoner. Therefore, a TBox classification was performed showing that Dione is consistent. As we stressed in [4], Dione is completed in the 93,3% of the classes. This means that the 93,3% of the total of ICD-10-CM classes have at least an axiom that defines that disease's category. The next objective in this research line is to provide a complete version of Dione by defining all classes with axioms. Therefore, we plan to add more mappings from other ontologies which are related to Dione. A first attempt is the work presented in [111] in which we implemented an algorithm that allows to populate axioms from mappings of ICD-10-CM and an target ontology from BioPortal. As a use case, we have used ORDO, which is an ontology that include rare diseases, genes and other features. The result was a more complete version of Dione with new axioms that define those ICD-10-CM diseases' categories mapped to the ORDO ones. According to these previous results, we are planning to replicate the same experiment for all ontologies mapped with ICD-10-CM to provide the most complete OWL version of ICD ever modeled.

Bibliography

[1] M. J. García-Godoy, I. Navas-Delgado, and J. Aldana-Montes. "Bioqueries: A Social Community Sharing Experiences While Querying Biological Linked Data". *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*. SWAT4LS '11. London, United Kingdom: ACM, 2012, pp. 24–31. ISBN: 978-1-4503-1076-5. DOI: 10.1145/2166896.2166906. URL: http://doi.acm.org/10.1145/2166896.2166906.

[2] M. J. García Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Sharing and executing linked data queries in a collaborative environment". *Bioinformatics* 29.13 (2013), pp. 1663–1670. DOI: 10.1093/bioinformatics/btt192.

[3] M. J. García-Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Reconstructing Hidden Semantic Data Models by Querying SPARQL Endpoints". *Database and Expert Systems Applications: 27th International Conference, DEXA 2016, Porto, Portugal, September 5-8, 2016, Proceedings, Part I*. Ed. by S. Hartmann and H. Ma. Cham: Springer International Publishing, 2016, pp. 405–415. ISBN: 978-3-319-44403-1. DOI: 10.1007/978-3-319-44403-1_25.

[4] M. del Mar Roldán-García, M. J. García-Godoy, and J. F. Aldana-Montes. "Dione: An OWL representation of ICD-10-CM for classifying patients' diseases". *Journal of Biomedical Semantics* 7.1 (2016). DOI: 10.1186/s13326-016-0105-x.

[5] R. S. Andreas Harth Katja Hose. *Linked Data Management*. Chapman and Hall/CRC, 2010. ISBN: 9781466582408.

[6] *The Open Linked Data Cloud*. http://lod-cloud.net/. Accessed: 28-March-2018.

[7] *Bio2RDF*. http://bio2rdf.org/sparql. Accessed: 6-July-2018.

[8] *Bio2RDF*. http://download.bio2rdf.org/. Accessed: 6-July-2018.

[9] *BioPortal*. http://sparql.bioontology.org/. Accessed: 6-July-2018.

[10] M. R. Kamdar and M. A. Musen. "PhLeGrA: Graph Analytics in Pharmacology over the Web of Life Sciences Linked Open Data". *WWW*. ACM, 2017, pp. 321–329.

[11] I. Navas-Delgado, M. J. García-Godoy, E. López-Camacho, M. Rybinski, A. Reyes-Palomares, M. Á. Medina, and J. F. Aldana-Montes. "kpath: integration of metabolic pathway linked data". *Database* 2015 (2015). DOI: 10.1093/database/bav053.

[12] A. Polleres, M. R. Kamdar, J. D. Fernández, T. Tudorache, and M. A. Musen. "A More Decentralized Vision for Linked Data". 2018.

[13] A. G. Nuzzolese, V. Presutti, A. Gangemi, A. Musetti, and P. Ciancarini. "Aemoo: Exploring Knowledge on the Web". *Proceedings of the 5th Annual ACM Web Science Conference*. WebSci '13. Paris, France: ACM, 2013, pp. 272–275. DOI: 10.1145/2464464.2464519.

[14] A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich. "Data Summaries for On-demand Queries over Linked Data". *Proceedings of the 19th International Conference on World Wide Web*. Raleigh, North Carolina, USA: ACM, 2010, pp. 411–420. DOI: 10.1145/1772690.1772733. URL: http://doi.acm.org/10.1145/1772690.1772733.

[15] C. Stadler, J. Lehmann, K. Höffner, and S. Auer. "LinkedGeoData: A Core for a Web of Spatial Open Data". *Semantic Web Journal* 3.4 (2012), pp. 333–354. URL: http://jens-lehmann.org/files/2012/linkedgeodata2.pdf.

[16] R. M. Carmona, A. Zafra, P. Seoane, A. J. Castro, D. Guerrero-Fernndez, T. Castillo-Castillo, A. Medina-García, F. M. Cánovas, J. Aldana-Montes, I. Navas-Delgado, J. D. D. Alché, and M. G. Claros. "ReprOlive: a Database with Linked Data for the Olive Tree (Olea europaea L.) Reproductive Transcriptome". *Frontiers in Plant Science* 6.625 (2015). DOI: 10.3389/fpls.2015.00625.

[17] V. Chelliah, N. Juty, I. Ajmera, R. Ali, M. Dumousseau, M. Glont, M. Hucka, G. Jalowicki, S. Keating, V. Knight-Schrijver, A. Lloret-Villas, K. N. Natarajan, J.-B. Pettit, N. Rodriguez, M. Schubert, S. M. Wimalaratne, Y. Zhao, H. Hermjakob, N. Le Novère, and C. Laibe. "BioModels: ten-year anniversary". *Nucleic Acids Research* 43.D1 (2015), pp. D542–D548. DOI: 10.1093/nar/gku1181.

[18] *T. Berners-Lee, "Linked Data", The World Wide Web Consortium (W3C), 27 07 2006.* http://www.w3.org/DesignIssues/LinkedData.html. Accessed: 28-July-2015.

[19] T. Berners-Lee, J. Hendler, and O. Lassila. "The Semantic Web". 284.5 (2001), pp. 34–43.

[20] *Linked Data principles by the W3C.* https://www.w3.org/wiki/LinkedData. Accessed: 1-Abril-2018.

[21] *RDF-Resource Data Framework.* https://www.w3.org/TR/rdf-concepts/. Accessed: 28-March-2018.

[22] W. Hu, H. Qiu, and M. Dumontier. "Link Analysis of Life Science Linked Data." *International Semantic Web Conference (2)*. Vol. 9367. Lecture Notes in Computer Science. Springer, 2015, pp. 446–462.

[23] *RDF.* https://www.w3.org/TR/rdf-concepts/. Accessed: 12-April-2018.

[24] *XML Extension Markup Language.* https://www.w3.org/TR/xml/. Accessed: 15-March-2018.

[25] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. "Bio2RDF: Towards a mashup to build bioinformatics knowledge systems". *Journal of Biomedical Informatics* 41.5 (2008), pp. 706 –716. DOI: https://doi.org/10.1016/j.jbi.2008.03.004.

[26] J. Hayes and C. Gutierrez. "Bipartite Graphs as Intermediate Model for RDF". *The Semantic Web – ISWC 2004*. Ed. by S. A. McIlraith, D. Plexousakis, and F. van Harmelen. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 47–61.

[27] *Data Set RDF Dumos.* https://www.w3.org/wiki/DataSetRDFDumps. Accessed: 15-March-2018.

[28] *SPARQL query language.* https://www.w3.org/TR/rdf-sparql-query/. Accessed: 15-March-2018.

[29] *Turle.* https://www.w3.org/TR/turtle/. Accessed: 15-March-2018.

[30] *N-triples.* https://www.w3.org/TR/n-triples/. Accessed: 15-March-2018.

[31] *RDF advantages.* https://www.w3.org/TR/WD-rdf-syntax-971002/. Accessed: 15-March-2018.

[32] *SPARQL version 1.1*. https://www.w3.org/TR/sparql11-query/. Accessed: 15-April-2018.

[33] *SPARQL version 1.0*. (https://www.w3.org/TR/rdf-sparql-protocol/). Accessed: 15-April-2018.

[34] M. Arenas, C. Gutierrez, and J. Pérez. "On the Semantics of SPARQL". *Semantic Web Information Management: A Model-Based Perspective*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 281–307. DOI: 10.1007/978-3-642-04329-1_13.

[35] *ARQ-An SPARQL processor for JENA*. https://jena.apache.org/documentation/query/. Accessed: 15-May-2018.

[36] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novère, H. Parkinson, E. Birney, and A. M. Jenkinson. "The EBI RDF platform: linked open data for the life sciences". *Bioinformatics* 30.9 (2014), pp. 1338–1339. DOI: 10.1093/bioinformatics/btt765. URL: http://dx.doi.org/10.1093/bioinformatics/btt765.

[37] *Web Ontology Language OWL*. https://www.w3.org/OWL/. Accessed: 15-May-2018.

[38] F. Baader. "Description Logics". *Reasoning Web. Semantic Technologies for Information Systems, 5th International Summer School 2009, Brixen-Bressanone, Italy, August 30 - September 4, 2009, Tutorial Lectures*. 2009, pp. 1–39. DOI: 10.1007/978-3-642-03754-2_1.

[39] *OWL2*. https://www.w3.org/TR/owl2-profiles/. Accessed: 15-May-2018.

[40] *OWL1*. https://www.w3.org/TR/2008/WD-owl11-syntax-20080108/. Accessed: 15-May-2018.

[41] B. Cuenca Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, and U. Sattler. "OWL 2: The next step for OWLCitation formats". *Web Semantics* 6.4 (Nov. 2008), pp. 309–322. DOI: 10.1016/j.websem.2008.05.001.

[42] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, and Z. Wang. "HermiT: An OWL 2 Reasoner". *J. Autom. Reason.* 53.3 (Oct. 2014), pp. 245–269. ISSN: 0168-7433. DOI: 10.1007/s10817-014-9305-1. URL: http://dx.doi.org/10.1007/s10817-014-9305-1.

[43] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, and Y. Katz. "Pellet: A Practical OWL-DL Reasoner". *Web Semant.* 5.2 (June 2007), pp. 51–53. DOI: 10.1016/j.websem.2007.03.004.

[44] F. Baader, C. Lutz, and B. Suntisrivaraporn. "CEL — A Polynomial-Time Reasoner for Life Science Ontologies". *Automated Reasoning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 287–291.

[45] E. Thomas, J. Z. Pan, and Y. Ren. "TrOWL: Tractable OWL 2 Reasoning Infrastructure". *Proceedings of the 7th International Conference on The Semantic Web: Research and Applications - Volume Part II*. Heraklion, Greece: Springer-Verlag, 2010, pp. 431–435. DOI: 10.1007/978-3-642-13489-0_38.

[46] Y. Kazakov, M. Krötzsch, and F. Simančík. "The Incredible ELK". *J. Autom. Reason.* 53.1 (2014), pp. 1–61. DOI: 10.1007/s10817-013-9296-3. URL: http://dx.doi.org/10.1007/s10817-013-9296-3.

[47] M. A. Musen. "The ProtÉGÉ Project: A Look Back and a Look Forward". *AI Matters* 1.4 (2015), pp. 4–12. ISSN: 2372-3483. DOI: 10.1145/2757001.2757003.

[48] *The NeOn Ontology Engineering Toolkit*. 2008. URL: http://watson.kmi.open.ac.uk/Downloads%20and%20Publications_files/neon-toolkit.pdf.

[49] K. Dentler, R. Cornet, A. ten Teije, and N. de Keizer. "Comparison of Reasoners for Large Ontologies in the OWL 2 EL Profile". *Semant. web* 2.2 (2011), pp. 71–87. DOI: 10.3233/SW-2011-0034.

[50] K. Donnelly. "SNOMED-CT: The advanced terminology and coding system for eHealth." *Stud Health Technol Inform* 121 (2006), pp. 279–90.

[51] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 2011. URL: http://linkeddatabook.com/.

[52] M. Papadaki, P. Papadakos, M. Mountantonakis, and Y. Tzitzikas. "An Interactive 3D Visualization for the LOD Cloud". *Proceedings of the Workshops of the EDBT/ICDT 2018 Joint Conference (EDBT/ICDT 2018), Vienna, Austria, March 26, 2018*. 2018, pp. 100–103.

[53] *Putting Government Data online*. https://www.w3.org/DesignIssues/GovData.html. Accessed: 3-July-2018.

[54] B. Villazón-Terrazas, L. M. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez. "Methodological Guidelines for Publishing Government Linked Data". *Linking Government Data*. Ed. by D. Wood. New York, NY: Springer New York, 2011, pp. 27–49. DOI: 10.1007/978-1-4614-1767-5_2.

[55] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. "DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia". *Semantic Web Journal* 6.2 (2015), pp. 167–195.

[56] *Wikipedia*. https://en.wikipedia.org/wiki/Wikipedia. Accessed: 2-July-2018.

[57] C. Stadler, J. Lehmann, K. Höffner, and S. Auer. "LinkedGeoData: A Core for a Web of Spatial Open Data". *Semantic Web Journal* 3.4 (2012), pp. 333–354.

[58] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. "Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections". *The Semantic Web: Research and Applications*. Ed. by L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 723–737.

[59] N. F. Noy, N. H. Shah, P. L. Whetzel, B. Dai, M. Dorf, N. Griffith, C. Jonquet, D. L. Rubin, M.-A. Storey, C. G. Chute, and M. a. Musen. "BioPortal: ontologies and integrated data resources at the click of a mouse." *Nucleic acids research* 37.Web Server issue (2009), W170–3. DOI: 10.1093/nar/gkp440.

[60] M. Salvadores, P. R. Alexander, M. A. Musen, and N. F. Noy. "BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF". *Semantic Web* 4.3 (2013), pp. 277–284. DOI: 10.3233/SW-2012-0086. URL: https://doi.org/10.3233/SW-2012-0086.

[61] M. Hausenblas. *Linked data application. DERI Technical Report*. Tech. rep. Accessed: 4-July-2018.

[62] D. A. Quan and R. Karger. "How to Make a Semantic Web Browser". *Proceedings of the 13th International Conference on World Wide Web*. New York, NY, USA: ACM, 2004, pp. 255–265. DOI: 10.1145/988672.988707.

[63] *Disco semantic web brwoser*. www4.wiwiss.fu-berlin.de/bizer/ng4j/disco. Accessed: 4-July-2018.

[64] L. Rutledge, J. van Ossenbruggen, and L. Hardman. "Making RDF Presentable: Integrated Global and Local Semantic Web Browsing". *Proceedings of the 14th International Conference on World Wide Web*. WWW '05. Chiba, Japan: ACM, 2005, pp. 199–206. ISBN: 1-59593-046-9. DOI: 10.1145/1060745.1060777. URL: http://doi.acm.org/10.1145/1060745.1060777.

[65] D. Huynh, S. Mazzocchi, and D. Karger. "Piggy Bank: Experience the Semantic Web Inside Your Web Browser". *The Semantic Web – ISWC 2005*. Ed. by Y. Gil, E. Motta, V. R. Benjamins, and M. A. Musen. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 413–430.

[66] S. Auer, R. Doehring, and S. Dietzold. "LESS - Template-Based Syndication and Presentation of Linked Data". *The Semantic Web: Research and Applications*. Ed. by L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 211–224.

[67] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets. "Tabulator: Exploring and Analyzing linked data on the Semantic Web". *Proceedings of the 3rd International Semantic Web User Interaction*. 2006.

[68] J. Koch and T. Franz. "LENA - Browsing RDF Data More Complex Than Foaf". *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008), Karlsruhe, Germany, October 28, 2008*. 2008.

[69] I. O. Popov, M. C. Schraefel, W. Hall, and N. Shadbolt. "Connecting the Dots: A Multi-pivot Approach to Data Exploration". *The Semantic Web – ISWC 2011*. Ed. by L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy, and E. Blomqvist. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 553–568.

[70] G. Kobilarov and I. Dickinson. "Humboldt: Exploring Linked Data". *Linked Data on the Web Workshop (LDOW2008) at WWW2008*. Beijing, China, 2008.

[71] M. Hildebrand, J. van Ossenbruggen, and L. Hardman. "/facet: A Browser for Heterogeneous Semantic Web Repositories". *The Semantic Web - ISWC 2006*. Ed. by I. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. M. Aroyo. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 272–285.

[72] P. Heim, J. Ziegler, and S. Lohmann. "gFacet: A Browser for the Web of Data". *Proceeding of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW'08)b*. 2008.

[73] S. F. C. Araújo, D. Schwabe, and S. D. J. Barbosa. "Experimenting with Explorator: a Direct Manipulation Generic RDF Browser and Querying Tool". *Visual Interfaces to the Social and the Semantic Web*. 2009.

[74] P. Haase, M. Schmidt, and A. Schwarte. "The Information Workbench As a Self-service Platform for Linked Data Applications". *Proceedings of the Second International Conference on Consuming Linked Data - Volume 782*. Bonn, Germany: CEUR-WS.org, 2010, pp. 119–124.

[75] *Marbles*. http://mes.github.io/marbles/. Accessed: 4-July-2018.

[76] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs. "Swoogle: A Search and Metadata Engine for the Semantic Web". *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*. CIKM '04. Washington, D.C., USA: ACM, 2004, pp. 652–659. DOI: 10.1145/1031171.1031289. URL: http://doi.acm.org/10.1145/1031171.1031289.

[77] G. Cheng, W. Ge, and Y. Qu. "Falcons: Searching and Browsing Entities on the Semantic Web". *Proceedings of the 17th International Conference on World Wide Web*. Beijing, China: ACM, 2008, pp. 1101–1102. DOI: 10.1145/1367497.1367676.

[78] G. Tummarello, R. Delbru, and E. Oren. "Sindice.Com: Weaving the Open Linked Data". *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*. ISWC'07/ASWC'07. Busan, Korea: Springer-Verlag, 2007, pp. 552–565.

[79] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, and S. Decker. "Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine". *Web Semantics: Science, Services and Agents on the World Wide Web* 9.4 (2011), pp. 365 –401. DOI: https://doi.org/10.1016/j.websem.2011.06.004.

[80] H. K. Azad, A. Deepak, and K. Abhishek. "Linked Open Data Search Engine". *Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies*. ICTCS '16. Udaipur, India, 2016, 17:1–17:5. DOI: 10.1145/2905055.2905075. URL: http://doi.acm.org/10.1145/2905055.2905075.

[81] *IBM Watson*. Accessed: 5-July-2018. URL: https://www.ibm.com/watson/.

[82] J. Weaver and P. Tarjan. "Facebook Linked Data via the Graph API". *Semant. web* 4.3 (2013), pp. 245–250.

[83] T. Steiner, R. Troncy, and M. Hausenblas. "How Google is using Linked Data Today and Vision For Tomorrow". *Proceedings of Linked Data in the Future Internet at the Future Internet Assembly (FIA 2010), Ghent, December 2010*. 2010.

[84] M. Samwald, A. Jentzsch, C. Bouton, C. S. Kallesøe, E. Willighagen, J. Hajagos, M. S. Marshall, E. Prud'hommeaux, O. Hassanzadeh, E. Pichler, and S. Stephens. "Linked open drug data for pharmaceutical research and development". *Journal of Cheminformatics* 3.1 (2011), p. 19. DOI: 10.1186/1758-2946-3-19.

[85] V. Momtchev, D. Peychev, T. Primov, and G. Georgiev. "Expanding the pathway and interaction knowledge in linked life data". *In Proc. of International Semantic Web Challenge*. 2009.

[86] E. Antezana, W. Blondé, M. Egaña, A. Rutherford, R. Stevens, B. De Baets, V. Mironov, and M. Kuiper. "BioGateway: a semantic systems biology tool for the life sciences". *BMC Bioinformatics* 10.10 (2009), S11. DOI: 10.1186/1471-2105-10-S10-S11.

[87] W. Fujibuchi, S. Goto, H. Migimatsu, I. Uchiyama, A. Ogiwara, Y. Akiyama, and M. Kanehisa. "DBGET/LinkDB: an integrated database retrieval system". *Pacific Symposium on Biocomputing* (1998), pp. 683–694.

[88] A. R. Kinjo, G.-J. Bekker, H. Suzuki, Y. Tsuchiya, T. Kawabata, Y. Ikegawa, and H. Nakamura. "Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures". *Nucleic Acids Research* 45.D1 (2017), pp. D282–D288. DOI: 10.1093/nar/gkw962.

[89] S. M. Wimalaratne, P. Grenon, H. Hermjakob, N. Le Novère, and C. Laibe. "BioModels linked dataset". *BMC Systems Biology* 8.1 (2014), p. 91. DOI: 10.1186/s12918-014-0091-5.

[90] R. Ranzinger, K. F. Aoki-Kinoshita, M. P. Campbell, S. Kawano, T. Lütteke, S. Okuda, D. Shinmachi, T. Shikanai, H. Sawaki, P. Toukach, M. Matsubara, I. Yamada, and H. Narimatsu. "GlycoRDF: an ontology to standardize glycomics data in RDF". *Bioinformatics* 31.6 (2015), pp. 919–925. DOI: 10.1093/bioinformatics/btu732.

[91] B. Bushman, D. Anderson, and G. Fu. "Transforming the Medical Subject Headings into Linked Data: Creating the Authorized Version of MeSH in RDF". *Journal of Library Metadata* 15.3-4 (2015), pp. 157–176. DOI: 10.1080/19386389.2015.1099967.

[92] J. Piñero, A. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong. "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants". *Nucleic Acids Research* 45.D1 (2017), pp. D833–D839. DOI: 10.1093/nar/gkw943.

[93] O. Bodenreider. *Lexical, terminological and ontological resources for biological text mining.* Artech House, 2006, pp. 43–66.

[94] A. Y. Wang, J. H. Sable, and K. A. Spackman. "The SNOMED clinical terms development process: refinement and analysis of content." *Proc AMIA Symp* (2002), pp. 845–849.

[95] C. Rosse and J. L. V. Mejino. "A reference ontology for biomedical informatics: the foundational model of anatomy". *Journal of Biomedical Informatics* 36.6 (2003), pp. 478–500. DOI: http://dx.doi.org/10.1016/j.jbi.2003.11.007.

[96] M. e. a. Ashburner. "Gene Ontology: Tool for the unification of biology". *Nature Genetics* 25 (2000), pp. 25–29. DOI: 10.1038/75556.

[97] D. Lindberg, B. Humphreys, and A. McCray. "The Unified Medical Language System". *Methods of Information in Medicine* 32.4 (1993), pp. 281–291.

[98] S. J. Nelson, D. Johnston, and B. L. Humphreys. *Relationships in Medical Subject Headings.* New York, NY, USA: Kluwer Academic Publishers, 2001, pp. 171–184.

[99] S Nitsuwat and W Paoin. "Development of ICD-10-TM Ontology for a Semi-automated Morbidity Coding System in Thailand." *Methods Inf Med* 51.5 (2012), 519—528. DOI: 10.3414/me11-02-0024.

[100] S. V. Pakhomov, J. D. Buntrock, and C. G. Chute. "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques." *Journal of the American Medical Informatics Association* 13.5 (2006), pp. 516–525. DOI: 10.1197/jamia.m2077.

[101] S. Pereira, A. Névéol, P. Massari, M. Joubert, and S. J. Darmoni. "Construction of a semi-automated ICD-10 coding help system to optimize medical and economic coding." *MIE*. Ed. by A. Hasman, R. Haux, J. van der Lei, E. D. Clercq, and F. H. R. France. Vol. 124. Studies in Health Technology and Informatics. IOS Press, 2006, pp. 845–850. DOI: 10.3233/978-1-58603-647-8-845.

[102] M. B. Sesen, E. Jiménez-Ruiz, R. Bañares-Alcántara, and M. Brady. "Evaluating OWL 2 Reasoners in the context of Clinical Decision Support in Lung Cancer Treatment Selection." *ORE*. Ed. by S. Bail, B. Glimm, R. S. Gonçalves, E. Jiménez-Ruiz, Y. Kazakov, N. Matentzoglu, and B. Parsia. Vol. 1015. CEUR Workshop Proceedings. CEUR-WS.org, 2013, pp. 121–127.

[103] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick. "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders". *Nucleic Acids Research* 30.1 (2002), pp. 52–55. DOI: 10.1093/nar/gki033.

[104] *ORPHANET website*. http://www.orpha.net/consor/cgi-bin/index.php. Accessed: 6-July-2018.

[105] M. Möller and S. Mukherjee. "Context-Driven Ontological Annotations in DICOM Images - Towards Semantic Pacs." *Proceedings of the Second International Conference on Health Informatics, HEALTHINF 2009, Porto, Portugal, January 14-17*. Ed. by L. Azevedo and A. R. Londral. INSTICC Press, May 20, 2009, pp. 294–299.

[106] G. Héja, G. Surján, G. Lukácsy, P. Pallinger, and M. Gergely. "GALEN based formal representation of ICD10." *I. J. Medical Informatics* 76.2-3 (2007), pp. 118–123. DOI: 10.1016/j.ijmedinf.2006.07.008.

[107] G. Héja, P. Varga, and G. Surján. "Design principles of DOLCE-based formal representation of ICD10." *MIE*. Ed. by S. K. Andersen, G. O. Klein, S. Schulz, and J. Aarts. Vol. 136. Studies in Health Technology and Informatics. IOS Press, 2008, pp. 821–826.

[108] M. Möller, D. Sonntag, and P. Ernst. "Modeling the International Classification of Diseases (ICD-10) in OWL". *Knowledge Discovery, Knowledge Engineering and Knowledge Management*. Vol. 272. 2013, pp. 226–240. DOI: 10.1007/978-3-642-29764-9_16. URL: http://dx.doi.org/10.1007/978-3-642-29764-9_16.

[109] I. Navas-Delgado, M. J. Garcia-Godoy, E. Lopez-Camacho, M. Rybinski, A. Reyes-Palomares, M. A. Medina, and J. F. Aldana-Montes. "kpath: integration of metabolic pathway linked data". *Database* 2015.0 (2015), bav053–bav053. DOI: 10.1093/database/bav053.

[110] M. J. García Godoy, E. López-Camacho, I. Navas-Delgado, and J. F. Aldana-Montes. "Bioqueries: a Social Community for SPARQL queries in Life Sciencess". *Actas de las XIX Jornadas de Ingeniería del Software y Bases de Datos. JISBD September 16-19*. 2014.

[111] M. J. García Godoy, E. López-Camacho, M. d. M. Roldán-García, and J. F. Aldana-Montes. "Enriquecimiento Automático de Ontologías Biomédicas mediante el uso de Mappings". *Actas de las XXIII Jornadas de Ingeniería del Software y Bases de Datos. JISBD September 17-19*. 2018.

[112] P. Heim, S. Hellmann, J. Lehmann, S. Lohmann, and T. Stegemann. "RelFinder: Revealing Relationships in RDF Knowledge Bases". *Semantic Multimedia*. Ed. by T.-S. Chua, Y. Kompatsiaris, B. Mérialdo, W. Haas, G. Thallinger, and W. Bailer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 182–187. ISBN: 978-3-642-10543-2.

[113] J. Brooke. "SUS-A quick and dirty usability scale". *Usability evaluation in industry* 189.194 (1996), pp. 4–7.

[114] *Hidden Semantic Model Algorithm*. https://github.com/estebanpua/ontology-endpoint-extraction. Accessed: 8-July-2018.

[115] *Hidden Semantic Model Web Service*. https://khaos.uma.es/oee. Accessed: 8-July-2018.

List of Figures